SCHOLARLY PUBLICATIONS

A CURRENT AWARENESS BULLETIN
OF RESEARCH OUTPUT

@DTU

(84th Edition)

DECEMBER 2019

BY: CENTRAL LIBRARY

DELHI TECHNOLOGICAL UNIVERSITY
(FORMERLY *DELHI COLLEGE OF ENGINEERING)*
GOVT. OF N.C.T. OF DELHI
SHAHBAD DAULATPUR, MAIN BAWANA ROAD
DELHI 110042

PREFACE

This is the **Eighty Fourth** Issue of Current Awareness Bulletin started by Delhi Technological University, Central Library. The aim of the bulletin is to compile, preserve and disseminate information published by the faculty, students and alumni for mutual benefits. The bulletin also aims to propagate the intellectual contribution of Delhi Technological University (DTU) as a whole to the academia.

The bulletin contains information resources available in the internet in the form of articles, reports, presentations published in international journals, websites, etc. by the faculty and students of DTU. The publications of faculty and student which are not covered in this bulletin may be because of the reason that the full text either was not accessible or could not be searched by the search engine used by the library for this purpose.

The learned faculty and students are requested to provide their uncovered publications to the library either through email or in CD, etc. to make the bulletin more comprehensive.

This issue contains the information published during **December, 2019**. The arrangement of the contents is alphabetical. The full text of the article which is either subscribed by the university or available in the web is provided in this bulletin.

Central Library

CONTENTS

1. A hybrid technique for simultaneous network reconfiguration and optimal placement of distributed generation resources, ***6.Imran Ahmad Quadri*** and ***3.S. Bhowmick***, Electrical, DTU

2. A New Method for Semantic Similarity Assessment using Fuzzy Formal Concept Analysis & Fuzzy Set Similarity Measure, Shivani Jain, Seeja.K.R and ***3.Rajni Jindal***, Computer Science & Engineering, DTU

3. A Review of The Combined Wind and Wave Energy Technologies, ***3.Dr. J.P. Kesari[1]***, ***8.Ashi Gupta[2]***, ***8.Kartik Shukla[2]***, ***8.Prachi Garg[2]***, [1]Mechanical and [2]CSE, DTU

4. A Review on Performance and Emission of CI Engine using Exhaust gas recirculation (EGR), ***7.Aseem Dubey***, ***3.Naushad Ahmad Ansari***, ***3.Girish Kumar***, ***3.Akhilesh Arora*** and Abhishek Sharma, Mechanical, DTU

5. A Study of Black Box Adversarial Attacks in Computer Vision, ***8.SIDDHANT BHAMBRI***, ***8.SUMANYU MUKU***, AVINASH TULASI and ARUN BALAJI BUDURU, Computer Science and Engineering, DTU

6. a-Bernstein–Kantorovich operators, ***3.Naokant Deo*** and ***6.Ram Pratap***, Applied Mathematics, DTU

7. Adaptive Weighted Graph Approach to Generate Multimodal Cancelable Biometric Templates, Gurjit Singh Walia, ***6.Gaurav Jain***, ***3.Nipun Bansal*** and Kuldeep Singh, Computer Science & Engineering, DTU

8. An ecological space based hybrid swarm-evolutionary algorithm for software reliability model parameter estimation, 6.Sangeeta, ***3.Kapil Sharma*** and Manju Bala, Computer Science and IT, DTU

9. An efficient copy move forgery detection using deep learning feature extraction and matching algorithm, ***3.Ritu Agarwal[1]*** and ***3.Om Prakash Verma[2]***, [1]TI and [2]Electronics and Communication, DTU

10. An impedimetric biosensor based on electrophoretically assembled ZnO nanorods and carboxylated graphene nanoflakes on an indium tin oxide electrode for detection of the DNA of Escherichia coli O157:H7, Nandita Jaiswal, ***6.Chandra Mouli Pandey1***, ***6.Shipra Solanki[2]***, Ida Tiwari and ***3.Bansi Dhar Malhotra[2]***, [1]Applied Chemistry and [2]Biotechnology, DTU

11. An Improved Authentication Protocol for Wireless Body Sensor Networks Applied in Healthcare Applications, ***6.1.Kakali Chatterjee***, CSE, DTU

12. Attention-guided deep convolutional neural networks for skin cancer classification, ***8.Arshiya Aggarwal***, ***8.Nisheet Das*** and ***6.Indu Sreedevi***, Electrical and Electronics, DTU

13. Carbon Nanotube Recessed Channel (CNT-RC) MOSFET for High Linearity/ULSI Applications, Ajay Kumar, Neha Gupta, ***6.Samarth Singh[1]***, ***6.Balark Tiwari[1]***, ***3.Madan Mohan Tripathi[2]*** and ***3.Rishu Chaujar[1]***, [1]Physics and [2]Electrical, DTU

14. Characterization and parametric optimization of % change in residual stress of Mo composite coating using Taguchi approach, ***6.Ankit Tyagi***, ***3.S.M Pandeya***, ***3.R S Walia***, ***3.Qasim Murtaza***, Mechanical, DTU

15. Comparative Study of Two Extensions of Heston Stochastic Volatility Model, ***6.Gifty Malhotra***, ***3.R. Srivastava*** and ***3.H.C. Taneja***, Applied Mathematics, DTU

16. Comparison of fuel characteristics of hydrotreated waste cooking oil with its biodiesel and fossil diesel, ***6.Ankit Sonthalia*** and ***3.Naveen Kumar***, Mechanical, DTU

17. Comparison of half effect absorption-compression cascaded refrigeration system using thermodynamic (energy-exergy) analysis, ***3.R.S. Mishra***, Mech & Production, DTU

18. Comparison of properties of ternary fuel blends of dieseloctanol with biodiesel, ***6.Sidharth*** and ***3.Naveen Kumar***, Mechanical, DTU

19. Computational fluid dynamics (CFD) modelling of hybrid photovoltaic thermal system, Ashish Saurabh, Deepali Atheaya and ***3.Anil Kumar***, Mechanical, DTU

20. Degradation of dyes using biologically synthesized silver and copper nanoparticles, ***7.Swati Raina***, ***7.Arpita Roy*** and ***3.Navneeta Bharadvaja***, Biotechnology, DTU

21. Development of Reliable and High Responsivity ZnO-Based UV-C Photodetector, Jitesh Agrawal , Tejendra Dixit , I. A. Palani, and ***8.1.Vipul Singh***, ECE, DTU

22. Development of the Ethyl Ester from Jatropa Oil through Response Surface Methodology Approach, ***3.Gautam, R***, ***3.Ansari, N***, Sharma, A. and Singh, Y, Mechanical, DTU

23. Digital Super Divide: Humanity on the Crossroad, ***8.1.K. K. Minocha***, DSM, DTU

24. Effect of Fiber Length and Content on Mechanical and Water Absorption Behavior of Coir Fiber-Epoxy Composite, ***6.Mohit Mittal*** and ***3.Rajiv Chaudhary***, Mechanical, Production, and Industrial, DTU

25. ELECTROCHEMO-MAGNETO ABRASIVE FLOW MACHINE SETUP FABRICATION AND EXPERIMENTAL INVESTIGATION OF THE PROCESS ALONGWITH MATHEMATICAL MODELING AND OPTIMIZATION, ***6.Sachin Dhull***, Ravinderjit Singh Walia, ***3.Qasim Murtaza*** and ***3.Mahendra Singh Niranjan***, Mechanical, DTU

26. EMISSION & PERFORMANCE ANALYSIS OF PRODUCER GAS AS IN IC ENGINE FUEL: A TECHNICAL REVIEW, Ankur Dixit, Abhishek Srivastava, Anmol, Akash Gupta, Nishant Garg and ***3.Rajiv Chaudhary***, Mechanical, DTU

27. Engineering properties of bacterially induced calcite formations, ***6.Sangeeta Shougrakpam*** and ***3.Ashutosh Trivedi***, Civil, DTU

28. Entropy as a measure of implied volatility in Options Market, ***3.H.C. Taneja***, ***6.Luckshay Batra*** and ***8.1.Pulkit Gaur***, Applied Mathematics, DTU

29. Fuzzy Logic Modelling of Degradation of Cement Mortar in Aggressive Environment, Kavita Verma, Shobha Ram, ***3.Alok Verma***, Civil, DTU

30. GPS Based Adaptive Traffic Light Timings and Lane Scheduling, ***6.Krishna Kant Sharma[1]*** and ***3.Indu S[2]***, [1]Electrical and [2]Electronics, DTU

31. GST IN INDIAN ECONOMY AND ITS POSSIBLE ROBUSTNESS, ***8.Mehul Gupta*** and ***8.Prashant Sharma***, Mechanical, Production and Automobile, DTU

32. Hiding Data in Images Using Cryptography and Deep Neural Network, Kartik Sharma, Ashutosh Aggarwal, ***8.Tanay Singhania***, Deepak Gupta and Ashish Khanna, Applied Mathematics, DTU

33. High Diversity Gain MIMO-Antenna for UWB Application with WLAN Notch Band Characteristic Including Human Interface Devices, ***8.Gaurav Saxena***, ***3.Priyanka Jain*** and Yogendra Kumar Awasthi, Electronics and Communication, DTU

34. Human action recognition using descriptor based on selective finite element analysis, ***3.Rajiv Kapoor[1]***, ***6.Om Mishra[1]*** and ***3.Madan Mohan Tripathi[2]***, [1]Electronics and [2]Electrical, DTU

35. Investigations on weld bead geometry and microstructure in CMT, MIG pulse synergic and MIG welding of AA6061-T6, ***6.Yashwant Koli***, ***3.N.Yuvaraj***, ***3.Vipin*** and S.Aravindan, Mechanical, DTU

36. Knowledge Discovery and Data Mining Review Papers, ***7.Zenawi Haileslassie Berhe*** and ***3.Vinod Kumar***, Computer, DTU

37. Land Use Land Cover Classification from Satellite Imagery using mUnet: A Modified Unet Architecture, ***7.Lakshya Garg***, Parul Shukla, Sandeep Kumar Singh, Vaishangi Bajpai and Utkarsh Yadav, Electrical and Electronics, DTU

38. Learn-able parameter guided Activation Functions, S.Balaji, ***8.T.Kavya*** and ***8.Natasha Sebastian***, Electrical and Electronics, DTU

39. Making Smart Homes Smarter: Optimizing Energy Consumption with Human in the Loop, ***8.Mudit Verma***, ***8.Siddhant Bhambri*** and Arun Balaji Buduru, Computer Science, DTU

40. Modelling Of Viscoelastic Fluid In Finishing Processes, ***3.Krovvidi Srinivas***, ***3.Q. Murtaza*** and ***3.A.K Agarwal***, Mechanical, DTU

41. Multifunction Filter/Inverse Filter Configuration Employing CMOS CDBAs, ***3.Ram Bhagat***[1], ***3.D. R. Bhaskar***[2], ***3.Pragati Kumar***[1], [1]Electrical and [2]Electronics, DTU

42. Multiple Resource Management and Burst Time Prediction using Deep Reinforcement Learning, ***8.Vaibhav Kumar***, ***8.Siddhant Bhambri***, ***3.Prashant Giridhar Shambharkar***, Computer, DTU

43. New holographic dark energy model with bulk viscosity in f(R, T) gravity, ***3.C P Singh*** and ***6.M Srivastava***, Applied Mathematics, DTU

44. On the resolution of the Sensitivity Conjecture, ***8.Rohan Karthikeyan***, ***8.Siddharth Sinha*** and ***8.Vallabh Patil***, Applied Mathematics, DTU

45. OPTIMIZATION OF ELECTRIC VEHICLE CHARGING MANAGEMENT WITH DISTRIBUTIONAL LOCATIONAL MARGINAL PRICING, ***8.1.Abhishek Tyagi*** and ***3.Ram Bhagat***, Electrical, DTU

46. Pricing of the Geometric Asian Options Under a Multifactor Stochastic Volatility Model, ***6.Gifty Malhotra***, ***3.R. Srivastava*** and ***3.H.C. Taneja***, A Mathematics, DTU

47. Probabilistic obstacle avoidance and object following: An overlap of Gaussians approach, Dhaivat Bhatt, ***8.Akash Garg***, Bharath Gopalakrishnan and K. Madhava Krishna, Electrical, DTU

48. Quality based adaptive score fusion approach for multimodal biometric system, ***6.Keshav Gupta***[1], Gurjit Singh Walia and ***3.Kapil Sharma***[2], [1]Computer and [2]IT, DTU

49. RATE OF CONVERGENCE OF GUPTA–SRIVASTAVA OPERATORS BASED ON CERTAIN PARAMETERS, ***6.RAM PRATAP*** AND ***3.NAOKANT DEO***, Applied Mathematics, DTU

50. Sentiment analysis using deep learning architectures: a review, ***6.Ashima Yadav*** and ***3.Dinesh Kumar Vishwakarma***, IT, DTU

51. Short Term GA-NFIS based Hybrid Method for Prediction of Wind Speed & Power in Sustainable Power Generation, Vijay Kumar, Yash Pal and ***3.Madan Mohan Tripathi***, Electrical, DTU

52. Skeleton based Activity Recognition by Fusing Part-wise Spatio-temporal and Attention Driven Residues, ***6.Chhavi Dhiman***, ***3.Dinesh Kumar Vishwakarma*** and ***6.Paras Aggarwal***, IT, DTU

53. Stacking Seismic Data Based on Ramanujan Sums, Karthikeyan Elumalai, Devendra Kumar Yadav, ***3.Anup Kumar Manpura*** and R. K. Patney, Electrical, DTU

54. Stochastic Gradient Descent–Whale Optimization Algorithm-Based Deep Convolutional Neural Network To Crowd Emotion Understanding, ***3.Avinash Ratre***, Electronics, DTU

55. Structural, Physical, Electrical and Dielectric Properties of Magnetic Glasses: xFe2O3 + (30 □ x)V2O5 + 30Na2O + 40B2O3 with x = 0 to 15, P. Chand, L. Kumar, ***6.A. Yadav*** and S. Khasa, Applied Physics, DTU

56. Studies on the effects of braking loads on a Railway Wheel, ***6.Aakash D Bansal***, N Ramachandran and ***3.V Rastogi***, Mechanical , DTU

57. Theoretical Analysis and Design of High-Stable-Gain Antenna with Ultrawide Band Capabilities and Suppressed Back Radiations, ***8.1.Sandeep Sharma***, Mainuddin, Binod Kumar Kanaujia and Mukesh Kumar Khandelwal, Electronics, DTU

58. Trap Assisted Persistent Photo-Conductivity in Solution-Processed CuO Thin Film, Akash Tripathi, Jitesh Agrawal, Tejendra Dixit and ***8.1.Vipul Singh***, Electronics, DTU

59. View-invariant Deep Architecture for Human Action Recognition using late fusion, ***6.Chhavi Dhiman*** and ***3.Dinesh Kumar Vishwakarma***, IT, DTU

60. Water Footprint Analysis of Delhi to Understand it's sustainability, ***7.Vaibhav***[1] and ***3.Dr Bharat Jhamnani***[2], [1]Environmental and [2]Civil, DTU

61. Wireless Sensor Network Based Controlling and Monitoring for Smart Homes Using Zig Bee, ***3.Bhavnesh Jaint1***, ***3.S. Indu***[2] and ***3.Neeta Pandey***[2], [1]Electrical and [2]Electronics, DTU

62. Xenic insurance company: ethical dilemma of an employee, ***3.Naval Garg***, DSM, DTU

1. *Chancellor*

2. *Pro Vice Chancellor*

2.1. *Ex Pro Vice Chancellor*

3. *Faculty*

3.1. *Ex Faculty*

4. *Teaching-cum-Research Fellow*

4.1. *Asst. Librarian*

5. *Alumni*

6. *Research Scholar*

6.1. *Ex Research Scholar*

7. *PG Scholar*

7.1. *Ex PG Scholar*

8. *Undergraduate Student*

8.1. *Ex Undergraduate Student*

Soft Computing
https://doi.org/10.1007/s00500-019-04597-w

METHODOLOGIES AND APPLICATION

A hybrid technique for simultaneous network reconfiguration and optimal placement of distributed generation resources

Imran Ahmad Quadri[1] · S. Bhowmick[1]

© Springer-Verlag GmbH Germany, part of Springer Nature 2019

Abstract
A new meta-heuristic method, comprehensive teaching learning harmony search optimization algorithm (CTLHSO), is developed in this paper for the simultaneous reconfiguration and optimal allocation of distributed generation resources in radial distribution systems. The proposed method is a hybridization of the teaching–learning-based optimization (TLBO) and the harmony search (HS) algorithms. Primarily, eleven mathematical benchmark functions are used to test the performance of the CTLHSO algorithm. The results are then compared with that of global best artificial bee colony (G-ABC), particle swarm artificial bee colony (PS-ABC), TLBO and improved TLBO (I-TLBO) with identical parameters and initial conditions. The results show that the CTLHSO performance is better than the G-ABC, PS-ABC, TLBO and I-TLBO. Subsequently, CTLHSO is implemented on the IEEE 33-bus and 69-bus radial distribution systems for network reconfiguration and optimal placement of distributed generation resources to minimize the power losses and improve the voltage profiles. Five case studies at three different load levels are carried out. The results obtained are found to be better than those obtained with HS algorithm, genetic algorithm, refined genetic algorithm and fireworks algorithm.

Keywords Distributed generation · Harmony search algorithm · Teaching–learning-based optimization · Network reconfiguration

Abbreviations

ASR	Auto-selection rate
bW	Bandwidth of each decision variable of solution vector
BTW	Best to worst
CTLHSO	Comprehensive teaching learning harmony search optimization
DG	Distributed generation
DN	Distribution network
DBW	Dynamic bandwidth
FWA	Fireworks algorithm
GA	Genetic algorithm
HM	Harmony memory
HMS	Harmony memory size
HSA	Harmony search algorithm
HMCR	Harmony memory consideration rate
$HMCR^{ci}$	HMCR value of the current iteration
PAR	Pitch adjustment rate
RDSs	Radial distribution systems
RGA	Refined genetic algorithm
StdDev	Standard deviation
SV	Solution vector
SVM	Solution vector matrix
TLBO	Teaching–learning-based optimization

Parameters

$HMCR_{min}$/ $HMCR_{max}$	Minimum/maximum value of HMCR
PAR_{min}/ PAR_{max}	Minimum/maximum value of PAR
$P_{L,k}/Q_{L,k}$	Real/reactive power demand of load at bus 'k'
IN	Number of improvisation
SVS	Total number of solution vectors in the SVM
AccRate	Improvisation acceptance rate percentage
AC_{th}	Threshold acceptance rate
n, nb	Total number of buses and branches in network, respectively

Communicated by V. Loia.

✉ Imran Ahmad Quadri
imranquadri83@gmail.com

[1] Electrical Engineering Department, Delhi Technological University, Delhi, India

Published online: 09 December 2019

nDG	Total number of DG in the network
BIBC	Bus injection to branch current matrix before reconfiguration
BIBC′	Bus injection to branch current matrix after reconfiguration
N	Total number of variables in solution vector
$X_{i,\min}/X_{i,\min}$	Minimum/maximum value of ith variable of solution vector in the SVM
$I_j^{\max}$	Magnitude of maximum loading in branch 'j' of the network
$V_{\min}/V_{\max}$	Minimum/maximum allowable voltage at all buses in the network
$P_{\mathrm{DG},k}^{\min}/P_{\mathrm{DG},k}^{\max}$	Minimum/maximum allowable real power supplied by DG in the network at bus 'k'
$Q_{\mathrm{DG},k}^{\min}/Q_{\mathrm{DG},k}^{\max}$	Minimum/maximum allowable reactive power supplied by the DG in the network at bus 'k'
R_j/X_j	Resistance/reactance of the line section 'j' between buses 'i' and 'k'

Variables

V_k and θ_k	Voltage amplitude and phase angle at bus 'k,' respectively
V_{m}	Minimum voltage in the network
m	Total number of buses beyond branch 'j' of the network
rand	Uniformly distributed random number between 0 and 1
I_j	Magnitude of current flowing through the branch 'j' of network
$\mathrm{SW}_{\mathrm{open}}$	Open tie/sectional switch in the network
X_i^{SV}	ith variable of SVth solution vector in SVM
$X_{\mathrm{old}}^{\mathrm{SV}}/X_{\mathrm{new}}^{\mathrm{SV}}$	SVth previous/generated solution vector in the SVM
$X_{\mathrm{old},i}^{\mathrm{SV}}/X_{\mathrm{new},i}^{\mathrm{SV}}$	ith variable of SVth previous/generated solution vector in the SVM
X_{teacher}	Solution vector in SVM having maximum fitness in each iteration
$X_{\mathrm{teacher},i}$	ith variable of solution vector in SVM having maximum fitness in each iteration
X_{worst}	Solution vector having worst fitness in the SVM
$X_{\mathrm{worst},i}$	ith element of solution vector having worst fitness in the SVM
X^{SV}	SVth solution vector in the SVM
X_i^{SV}	ith variable of SVth solution vector in the SVM
$P_{\mathrm{grid}}/Q_{\mathrm{grid}}$	Active/reactive power supplied by grid to the network
$P_{\mathrm{loss}}/Q_{\mathrm{loss}}$	Total network active/reactive power loss before network reconfiguration and DG allocation
$P_{\mathrm{loss}}^r/Q_{\mathrm{loss}}^r$	Total network active/reactive power loss after reconfiguration
$P_{\mathrm{loss}}^{\mathrm{DG}}/Q_{\mathrm{loss}}^{\mathrm{DG}}$	Total active/reactive power loss in the network after DG placement
$P_{\mathrm{loss}}^{\mathrm{DG},r}/P_{\mathrm{loss}}^{\mathrm{DG},r}$	Total network active/reactive power loss after DG placement and reconfiguration
I_j	Magnitude of current flowing through the branch 'j' of network
M_i	Mean of ith decision variable of solution vector
T_{F}	Teaching factor
a	Teaching factor rate
A	Incidence matrix of the distribution network
H	Load level of the distribution network
$F\left(X_{\mathrm{old}}^{\mathrm{SV}}\right)$	Fitness of previous SVth solution vector in SVM

1 Introduction

Power loss, voltage deviation, stability, operational costs and environmental issues are the major concerns of distribution network operators. Optimal DG allocation along with the network reconfiguration in distribution systems significantly eases the challenges of distribution networks. So, proper optimization techniques must be used to optimally allocate the DGs and reconfigure the distribution network by improving the network operating conditions.

Several comprehensive research works have been reported in the area of optimal allocation of DGs to improve the distribution system performance. A non-iterative analytical method for single DG placement in both radial and meshed systems has been presented in Wang and Nehrir (2004) to minimize system power losses for both time-varying and time-invariant loads in the IEEE 6-bus and 30-bus distribution systems. A very efficient analytical method employing BIBC approach for the siting and sizing of DGs in the IEEE 12-bus, 34-bus and 69-bus system has been proposed in Gözel and Hocaoglu (2009). However, Gözel and Hocaoglu (2009) limited the placement of a single DG. An analytical method for multiple DG placements based on the active and reactive components of the branch currents has been reported in Gopiya Naik et al. (2015). Hung et al. (2010) have presented an improved analytical (IA) method for single DG placement and have highlighted its superiority over existing methods based on the assumption that the DG power factor is same as the

aggregate load power factor. This method reduces the computational time significantly while considering various types of DGs to minimize power losses and improve the voltage profile of the network. Although analytical methods for siting and sizing of DGs have fast convergence, however, with the increase in the type and number of DGs, the computational complexity increases and may lead to non-optimal solutions.

Advances in soft computing techniques have led to the development of several optimization algorithms for the optimal allocation of DGs. In this perspective, Singh et al. (2008), Ameli et al. (2014), Gomez-Gonzalez et al. (2012), Moradi and Abedini (2012), Niknam et al. (2011), Saha and Mukherjee (2016), Abu-Mouti and El-Hawary (2011), Sheng et al. 2016), Mohanty and Tripathy (2016) and Imran et al. (2014) present some comprehensive research works on the optimal allocation of type-1 DGs. Singh et al. (2008) have carried out the DG placement using genetic algorithm (GA) in the IEEE 16-bus, 37-bus and 75-bus radial distribution networks with different loading conditions. However, GA requires more computational time while suffering from premature convergence. Multi-objective particle swarm optimization (PSO) for DG placement in the IEEE 33-bus radial distribution system (RDS) has been proposed by Ameli et al. (2014) for power loss reduction, improvement of voltage profile and stability. Gomez-Gonzalez et al. (2012) have employed discrete PSO for optimal DG allocation to reduce power losses in the IEEE 30-bus radial distribution network. A hybrid GA-PSO-based algorithm for optimal DG allocation to reduce power losses and improve voltage profile and stability has been presented in Moradi and Abedini (2012). A multi-objective modified honey bee mating optimization algorithm for siting and sizing of DG has been reported in Niknam et al. (2011). The proposed method avoids strong local convergence as compared to the original honey bee mating optimization algorithm. Several nature-inspired algorithms have also been proposed by researchers for the siting and sizing of DGs. Some notable ones include the chaotic symbiotic organism search (CSOS) algorithm (Saha and Mukherjee 2016) and the artificial bee colony (ABC) algorithm (Abu-Mouti and El-Hawary 2011). Sheng et al. (2016) have implemented multi-objective harmony search algorithm (HSA) for optimal DG allocation in the 33-bus RDS. TLBO method for the optimal siting and sizing of DGs has been reported in Mohanty and Tripathy (2016). Imran et al. (2014) have presented a new swarm intelligence-based FWA for simultaneous reconfiguration and DG placement in distribution systems. A multi-objective, modified teaching–learning-based optimization algorithm for the optimal allocation of DGs in the 33-bus, 69-bus and 118-bus RDS has been presented in Martín García and Gil Mena (2013). Although this algorithm has fast convergence but gives non-optimal solution when the number of solution variables increases. Hence, an algorithm is required that can provide an optimal solution when variables are more as well as having fast convergence rate.

1.1 Motivation and aim

Although analytical methods have no convergence issues, with an increase in the number and type of DGs, their complexity increases, with a consequent increase in the computational time. This is particularly true for multi-objective formulations with a large number of equality and inequality constraints. Many other optimization techniques (Geem and Kim 2018; Song 2013; Ali 2012; Kalajdzic et al. 2015; Al-Janabi 2018; Al-Janabi et al. 2015, 2018; Al-Janabi and Alkaim 2019; Al-Janabi and Ibrahim 2017) exist, but they require proper parameter tuning for fast convergence. Recently, TLBO method (Rao et al. 2011) has been developed, which is almost independent of the optimization parameters and possesses excellent convergence characteristics for lesser number of variables. However, it has the tendency to be trapped in strong local minima when the number of variables increases. A modified TLBO algorithm for DG placement has been reported in Mohanty and Tripathy (2016) where the additional mutation phase is considered for finding the global solution. Sultana and Roy (2014) have presented a quasi-oppositional TLBO (QOTLBO) for DG placement in RDS. However, these features add complexity to the TLBO and increases the computational time. An improved TLBO has been reported in Kanwar et al. (2015) in which a crossover rate and a crossover parameter have to be specified. However, both of them need to be tuned to achieve a satisfactory convergence. Harmony search (HS) is another swarm intelligence algorithm, which shows good exploration capability but poor exploitation capability for global solutions. Several HS variants have been suggested to solve complex optimization problems, such as self-adaptive global harmony search algorithm (SGHS) (Pan et al. 2010), IHS (Mahdavi et al. 2007), intelligent tuned harmony search algorithm (ITHS) (Yadav et al. 2012), enhance harmony search (EHS) (Das et al. 2011), novel global harmony search (NGHS) (Zou et al. 2010), DIHS (Kattan and Abdullah 2013), NDHS (Tuo et al. 2015) and DSHS (Chen et al. 2012). However, these improved versions of HS are also unable to handle complex optimization problems of high dimensionality and modality. IHS lacks in precise solution. It is observed that NGHS, SGHS and NDHS get easily trapped into strong local minima/maxima.

Although the solution provided by EHS and DSHS is satisfactory, they require more convergence time for the global solution with high-dimensional problems. It is observed that TLBO exhibits a good search space exploitation capability, while HS has a good exploration capability. So, a proper integration of the merits of TLBO and HS would result in a better optimization technique for high-dimensional and multimodality problems.

To get a good balance of the exploration and exploitation of the search space, a new optimization algorithm based on the hybridization of HS and TLBO is presented in this paper. In this algorithm, at first, the HS algorithm is utilized to explore the search space with a high probability of finding the global solution. Subsequently, TLBO carries out the exploitation of the search space for the global solution. For choice of TLBO or HS, the proposed algorithm uses self-adaptive selection probability.

In this paper, CTLHSO has been proposed for the simultaneous reconfiguration and optimal allocation of DGs in distribution networks. The proposed technique is a hybrid of the HSA and the TLBO. While the HSA is more proficient in the exploration of the search space, the TLBO algorithm is more adept at its exploitation. The CTLHSO algorithm extracts the best qualities of both the HSA and TLBO, which makes it faster and immune to local extrema trappings. The proposed algorithm is primarily tested on 11 mathematical benchmark functions for its validation. Subsequently, the proposed algorithm is implemented on the IEEE 33-bus and 69-bus radial distribution systems. Results validate the superiority of the proposed method over FWA, GA, RGA and HSA. A comparison of the reduction in power losses and voltage profile improvement without and with DGs in the 33- and 69-bus radial distribution systems before and after reconfiguration shows improvement over HSA, GA, RGA and FWA algorithms (Imran et al. 2014; Tuo et al. 2015). These reiterate the superiority of the proposed method in the perspective of computational speed, accuracy and immunity to local extrema trappings, over existing HSA, GA, RGA and FWA algorithms.

1.2 Research contribution

The contribution of the paper can be summarized as follows:

- Development of a CTLHSO algorithm which is capable of handling mixing integer variables and constrained optimization problems.
- Modification in the teaching phase improves the exploitation and exploration capability of the proposed algorithm over HSA and TLBO, which improves immunity to local extrema trappings and ensures a global solution.
- Direct power flow method based on the BIBC/BCBV matrix is used for the power flow which does not require either the network admittance matrix or the forward/backward substitution of the Jacobian matrix. This reduces the computational time which is substantially less than exact loss formulation (Das et al. 2011).
- Optimal number of DGs is selected for placement in the IEEE 33- and 69-bus distribution systems for the maximum rate of reduction in power loss.
- Multiple case studies were conducted with standard mathematical benchmark functions and optimal allocation of DGs without and with reconfiguration in the IEEE 33-bus and 69-bus radial distribution test systems.
- Results demonstrate a marked improvement in the performance indices (network active power loss, voltage profile improvement, etc.) of the distribution test systems over HSA, GA, RGA and FWA.

The rest of the paper is organized as follows: Sect. 2 presents the introduction and the proposed modifications carried out in TLBO and HSA. Section 3 illustrates the implementation of the proposed technique. Section 4 deals with the problem formulation. Section 5 deals with the mathematical validation of the proposed algorithm and its implementation in several radial distribution systems. Section 6 presents the conclusion and future scope of work.

2 Outline of the algorithms

2.1 Teaching–learning-based optimization

TLBO algorithm was first introduced in Rao et al. (2011). TLBO is a nature-inspired algorithm where the best learner, i.e., teacher, improves the performance of the other remaining learners known as the '*teaching phase*' and each learner improves his knowledge by interacting with other fellow learners in the '*learning phase*.' In this manner, with proper interaction, the TLBO proceeds toward the global solution.

The steps of TLBO are as follows:

Step 1 Initialize the solution vector matrix and algorithm parameters.
Initially, SVM is initialized where each row represents a solution vector (X^{SV} where SV = 1, 2, 3…SVS). Each decision variable (X_i^{SV}) of SVth solution vector is generated using Eq. (1a) if the decision variable is continuous number or Eq. (1b) if the decision variable is discrete number as shown below.
If X_i^{SV} is continuous number

$$X_i^{SV} = X_{i,\min} + \text{rand} * \left(X_{i,\min} - X_{i,\max}\right) \quad \text{where } i = 1, 2, 3\ldots N \tag{1a}$$

If X_i^{SV} is discrete number

$$X_i^{SV} = \text{round}\left(X_{i,\min} + \text{rand} * \left(X_{i,\min} - X_{i,\max}\right)\right) \quad \text{where } i = 1, 2, 3\ldots N \tag{1b}$$

The initialized solution vectors are stored in SVM as shown below,

$$\text{SVM} = \begin{bmatrix} X_1^1 & X_2^1 & X_3^1 & \ldots & X_{N-1}^1 & X_N^1 \\ X_1^2 & X_2^2 & X_3^2 & & X_{N-1}^2 & X_N^2 \\ \vdots & \vdots & \vdots & \ldots & \vdots & \vdots \\ X_1^{SVS-1} & X_2^{SVS-1} & X_3^{SVS-1} & \ldots & X_{N-1}^{SVS-1} & X_N^{SVS-1} \\ X_1^{SVS} & X_2^{SVS} & X_3^{SVS} & & X_{N-1}^{SVS} & X_N^{SVS} \end{bmatrix} \tag{2}$$

The decision variable (X_i^{SV}) of solution vector (X^{SV}), which is continuous, will be generated by Eq. (1a), and the discrete decision variable will be generated by Eq. (1b). However, it can be observed that Eq. (1a) is a generating continuous decision variable $\left(X_i^{SV}\right)$, but when X_i^{SV} is rounded off to get the discrete number, the result of Eq. (1b) will be the same. So for our convenience now in this paper, both continuous and discrete decision variable generate only one equation which will be given.

Step 2 Generate a new solution vector.

A. Teacher phase: The teacher will continuously be augmenting the mean marks (knowledge) of a particular course (variable) to the best of his capacity of the students of the class. So, a random process must take place in order to improve the knowledge of students of the class. For each student, the marks/decision variable is generated detailed in Eq. (3) as shown below:

$$X_{\text{new},i}^{SV} = X_{\text{old},i}^{SV} + \text{rand} * \left(X_{\text{Teacher},i} - T_F M_i\right). \tag{3}$$

T_F can be either 1 or 2 (Al-Janabi et al. 2015). The new generated solution vector X_{new}^{SV} is accepted if its fitness (marks) is better than that corresponding to the X_{old}^{SV}.

B. Learner phase: This is an alternative way to improve the knowledge by one's own effort without the teacher. Students randomly interact with other students of the class and improve their understanding of a particular subject. The learner phase is mathematically expressed by Eq. (4a) and (4b) as,

$$X_{\text{new},i}^{SV} = X_{\text{old},i}^{SV} + \text{rand} * \left(X_{\text{old},i}^{p} - X_{\text{old},i}^{q}\right) \quad \text{if } F\left(X_{\text{old}}^{p}\right) < F\left(X_{\text{old}}^{q}\right) \tag{4a}$$

$$X_{\text{new},i}^{SV} = X_{\text{old},i}^{SV} + \text{rand} * \left(X_{\text{old},i}^{q} - X_{\text{old},i}^{p}\right) \quad \text{if } F\left(X_{\text{old}}^{p}\right) > F\left(X_{\text{old}}^{q}\right), \tag{4b}$$

where 'SV,' '*p*,' 'q' are the learners in the class in such a way that SV $\neq p \neq q$ and $F(X_{\text{old}}^{p})$ represents the fitness (marks) of a SVth learner in a class.

Step 3 Update the solution vector matrix.
If the fitness corresponding to 'X_{new}^{SV}' is found to be better than 'X_{old}^{SV},' 'X_{new}^{SV}' solution vector is accepted for both teaching and learning phases, and SVM is updated otherwise it is rejected.

Step 4 Check the termination criterion.
The TLBO keeps on repeating the steps 1, 2 and 3, till a maximum number of improvisation (*IN*) are met, which is set during the parameter initialization process.

So, these two stages, i.e., *teaching* and *learning phases*, comprise iteration. After several iterations, the global

solution is reached by the TLBO algorithm. Although this algorithm has the minimum number of parameters to be tuned as compared to other population-based optimization techniques, it suffers from premature convergence due to strong local minima trappings of the objective function.

2.2 Harmony search algorithm

HSA (Pan et al. 2010; Mahdavi et al. 2007; Yadav et al. 2012; Das et al. 2011; Zou et al. 2010; Kattan and Abdullah 2013; Tuo et al. 2015; Chen et al. 2012) mimics the improvisation of music players where the sounds for better aesthetic estimation improved through practice after practice. Similarly, values of objective function can be improved iteration by iteration. The HSA can be realized considering the following steps:

Step 1 Initialize the HM and algorithm parameters.
Initialization of HM matrix, which is same as SVM initialization in TLBO. Other parameters are HMS and HMCR, which represent the probability for the new value of solution vector for HM; PAR represents the probability to shift decision variable to neighboring value within the possible range; NI or stopping criterion

Step 2 Improvisation of solution vectors
A new set of solution vectors are generated based on three parameters: (1) HMCR, (2) PAR and (3) random selection. Generation of new solution vector of HM is known as improvisation. HMCR varies between 0 and 1, which is the probability of choosing the decision variable value from the HM, while (1-HMCR) is the probability of selecting the variable from the possible range of values, as given below

$$\begin{aligned}&\text{if} \quad \text{rand} < \text{HMCR}\\ &\quad X_{\text{new}}^{\text{SV}} \leftarrow X_{\text{old}}^{\text{SV}}\\ &\quad\quad \text{if rand} < \text{PAR}\\ &\quad X_{\text{new}}^{\text{SV}} \leftarrow X_{\text{old}}^{\text{SV}} \pm \text{rand} * \text{bW}\\ &\quad\quad \text{end}\\ &\text{else}\\ &\quad X_{\text{new}}^{\text{SV}} \text{calculated using Eq.}(1)\\ &\text{end}\end{aligned} \tag{5}$$

The old solution vector is retained when it satisfies HMCR condition only. However, new solution vector is generated as per Eq. (5) when it satisfies the HMCR and PAR conditions, otherwise according to Eq. (1). This step is decided by PAR, where bW is arbitrarily chosen values lying between 0 and 1.

Step 3 Update harmony memory
If the value of $F(X_{\text{new}}^{\text{SV}})$ is better than $F(X_{\text{worst}}^{\text{SV}})$, the new solution vector $(X_{\text{new}}^{\text{SV}})$, the new generated $X_{\text{new}}^{\text{SV}}$ is accepted otherwise rejected.

Step 4 Algorithm termination criterion.
The HSA kept on repeating the steps 1, 2 and 3, till a *IN* is met, which is set during the parameter initialization process.

2.3 Modification in TLBO and HS algorithms

There are several modifications proposed in TLBO and HSA that the individual algorithm has better exploration and exploitation capabilities before hybridization of TLBO and HSA.

2.3.1 Modification in TLBO

The following modifications are proposed to improve the performance of the TLBO and avoid premature convergence:

2.3.1.1 Modification in teaching phase In conventional TLBO, the new solution vector is generated based on the mean of the particular decision variable. In the proposed technique, instead of the mean, the worst solution vector of the SVM (i.e., having the worst fitness in the SVM) is selected. This results in a wider search space for the new generated solution vector '$X_{\text{new}}^{\text{SV}}$' and a higher probability of reaching the global optimized solution.

Thus, the modification achieved is shown below as:

$$X_{\text{new},i}^{\text{SV}} = X_{\text{old},i}^{\text{SV}} + rand * \left(X_{\text{Teacher},i} - T_F * X_{\text{worst},i}\right) \tag{6}$$

where $X_{\text{worst},i}$ is the ith decision variable of worst solution vector (X_{worst}), having the worst fitness in the SVM.

2.3.1.2 Modification in the teaching factor (T_F) TLBO method considers 'T_F' either 1 or 2, which corresponds to a transfer of 0% or 100% knowledge from the teacher to the learner, respectively. But practically this assumption is incorrect as it should be between 0 and 100%. So this T_F is modified as per Eq. (7), given below as:

$$T_F = \left(1/\text{rand}\right)^a \tag{7}$$

where 'a' is defined as the teaching factor rate. If the value of 'a' is high, it increases the search space and hence the probability of reaching the global solution. It is observed that keeping the value of 'a' between 0 and 5 yields better results with several objective functions, as highlighted in the case studies and results. So, this modification results in

an improved transfer of knowledge than conventional TLBO.

2.3.2 Modification in HSA

In basic HS (Geem et al. 2001), parameters of HMCR, PAR and bW are constant values. Our proposed algorithm adopts dynamic strategies to select the value of HMCR and PAR as same as Geem et al. (2001). Dynamic characteristics of HMCR and PAR provide better balance in exploration and exploitation of search space in this hybrid algorithm.

$$\mathrm{HMCR}^{\mathrm{ci}} = \mathrm{HMCR}_{\max} - (\mathrm{HMCR}_{\max} - \mathrm{HMCR}_{\min}) * \left(\frac{\mathrm{currentiteration}}{\mathrm{IN}}\right)^{s} \tag{8}$$

where 's' is exponent and can be the integer between 1 and 3. currentiteration is the ongoing iteration of optimization process.

2.3.2.1 Dynamic bW and PAR During each iteration of the algorithm, HM contains several solution vectors, which consists of best and worst solution vectors determined by their fitness values. The variation of PAR and bW is the function of the ongoing iteration and IN. So, the proper selection of a number of IN is needed for faster convergence of solution vector. A new approach has been adopted in Kattan and Abdullah (2013) to tune the value of bW and PAR, which is selected on the basis of quality (fitness) of the current solutions. The dynamic value of bW and PAR is based on the best-to-worst (BTW) ratio given by Eq. (9). BTW is always less than 1, for minimization problems, while for maximization problems BTW is obtained by minimizing the inverse of the fitness function values. So, the value of BTW will be always high, which will ensure that the current SVM solution vectors are converging toward the current best solution vectors. When the value of BTW decreases, PAR value will increase, which will ensure better exploitation ability of the algorithm, while the increment in the value of BTW will result in lower PAR value, improving the exploration capability of the optimization algorithm. PAR and DBW for the decision variables are calculated using Eqs. (10) and (13), using T and StdDev of the respective SVM column, instead of the entire population. T is self-adaptive in nature, and initially, its value is set to 2.0. When $T > 1$, it extends the range of the dynamic bandwidth to explore the search space outside the current boundary (lies between the minimum and maximum), while the lower value of K causes convergence of algorithm to a better solution. The AccRate is the total number of accepted improvisations in the last 100 iterations. $\mathrm{AC_{th}}$ is a threshold at which the value of T is decreased in proportion to the AccRate. Experimental analysis shows that $\mathrm{AC_{th}} = 0.02$produces better solutions. When the current improvisation acceptance rate falls below $\mathrm{AC_{th}}$, the range of bW begins to decrease linearly, which helps in fine-tuning of the decision variable.

$$\mathrm{BTW} = \frac{F(X_{\mathrm{teacher}})}{F(X_{\mathrm{worst}})} \tag{9}$$

$$\mathrm{PAR}^{\mathrm{Mod}} = \mathrm{PAR}_{\max} + \mathrm{BTW} * (PAR_{\min} - \mathrm{PAR}_{max}) \tag{10}$$

$$T = \begin{Bmatrix} 2 & \mathrm{AccRate} \geq \mathrm{AC_{th}} \\ 2 * \dfrac{\mathrm{AccRate}}{\mathrm{AC_{th}}} & \mathrm{AC_{th}} > \mathrm{AccRate} > 1 \\ 0.1 & \mathrm{AccRate} \leq 1 \end{Bmatrix} \tag{11}$$

$$\mathrm{bW}(i) = T * \mathrm{StdDev}\left(X^{\mathrm{SV}}_{\mathrm{old},i}\right) \tag{12}$$

$$\mathrm{DBW}(i) = \mathrm{rand}(-\mathrm{bW}(i), \mathrm{bW}(i)) \tag{13}$$

After modification in HMCR and PAR in the HSA, the new solution vector $(X^{\mathrm{SV}}_{\mathrm{new}})$ will be generated, as given below:

$$\begin{aligned} &\text{if} \quad \mathrm{rand} < \mathrm{HMCR}^{\mathrm{ci}} \\ &\quad X^{\mathrm{SV}}_{\mathrm{new}} \leftarrow X^{\mathrm{SV}}_{\mathrm{old}} \\ &\quad\quad \text{if} \quad \mathrm{rand} < \mathrm{PAR}^{\mathrm{Mod}} \\ &\quad\quad\quad X^{\mathrm{SV}}_{\mathrm{new}} \leftarrow X^{\mathrm{SV}}_{\mathrm{old}} \pm \mathrm{rand} * \mathrm{DBW} \\ &\quad\quad \text{end} \\ &\text{else} \\ &\quad X^{\mathrm{SV}}_{\mathrm{new}} \quad \text{calculated using Eq. (1)} \\ &\text{end} \end{aligned} \tag{14}$$

2.4 Proposed CTLHSO algorithm

It is observed that the algorithm characteristics of TLBO and HS are complimentary to each other, i.e., HSA is having excellent exploration behavior but slow in convergence, while TLBO exhibits best exploitation characteristics and fast convergence. Therefore, our aim in this proposed algorithm is to utilize merits of both algorithms and find the global solution with fast convergence. For the maximum exploration of the search space initially, HS is employed and then exploitation is done with the help of TLBO. Hence, in this new approach, selection of teaching phase of TLBO and local pitch adjustment by HMCR of HSA is based on the ASR as shown in Eq. (15).

$$\mathrm{ASR} = \frac{F\left(X^{\mathrm{SV}}_{\mathrm{teacher}}\right)}{F\left(X^{\mathrm{SV}}_{\mathrm{worst}}\right)} \tag{15}$$

For minimization problems, the value of best fitness will always less than worst fitness value, so ASR will be less than 1, while for maximization problems, ASR is obtained by minimizing the inverse of the fitness function values, which will also lead to ASR less than 1. Therefore, ASR is

always less than 1. ASR is dynamic in nature and self-adaptive. Initially, ASR will be low that will give preference to the HS algorithm, which will explore more search space, and when ASR reaches toward unity, teaching phase is selected for better tuning of the variables.

The CTLHSO works as follows:

To minimize the demerits of HS and TLBO, the CTLHSO initially utilizes the exploration characteristics of HS, and then, TLBO algorithm is implemented for fast convergence of solution vectors. The following steps are considered for CTLHSO algorithm.

A. Solution vector selection: $X_{\text{new}}^{\text{SV}} = X_{\text{old}}^{\text{SV}}$ is solution vector from SVM.

B. Generation of solution vector $(X_{\text{new}}^{\text{SV}})$ in each iteration comprises four steps.

Step 1 Auto-selection rate (ASR).
The selection of teaching phase or HMCR for the generation of target vector X_{new} depends upon the ASR as given in Eq. (15).

Step 2 Teaching phase/harmony memory consideration rate and local pitch adjustment/mutation phase
When ASR is less than '*r*' where r is the number between 0 and 1, teaching phase is selected, and Eq. (16) is used to generate $X_{\text{new}}^{\text{SV}}$. If ASR is more than *r*, HSA is selected for the generation of $X_{\text{new}}^{\text{SV}}$. The probability of tuning the target vector is decided by the HMCR. If the random number is less than HMCR, local pitch adjustment of $X_{\text{new}}^{\text{SV}}$ is carried out by using Eq. (14); otherwise, mutation phase is carried out according to Eq. (1) to generate new target vector $X_{\text{new}}^{\text{SV}}$ in feasible space. If the fitness of target vector $X_{\text{new}}^{\text{SV}}$ is better than the fitness of worst fitness vector, then X_{worst} is replaced by $X_{\text{new}}^{\text{SV}}$.

Step 3 Learning phase.
This phase enhances the knowledge (fitness) of target vector $X_{\text{new}}^{\text{SV}}$ by interacting with the randomly selected harmony vector from harmony memory. Improvement of $X_{\text{new}}^{\text{SV}}$ is done according to Eq. (4a) and (4b). If the fitness of target vector $X_{\text{new}}^{\text{SV}}$ is better than the fitness of worst fitness vector, then X_{worst} is replaced by $X_{\text{new}}^{\text{SV}}$.

Step 4 Dynamic adjustment of parameters.
For efficient operation of CTLHSO algorithm, the HMCR^{ci}, PAR^{Mod} and DBW parameters are dynamically tuned, in each iteration as per Eqs. (8), (10) and (13).

C. If the new solution vectors generated have better fitness value, then update the SVM, otherwise retain the old solution vector and go for the next iteration. If a number of iteration satisfy the condition for termination, terminate the algorithm and best solution vector of HM is selected as the solution of the objective function.

The above steps of the proposed algorithm, i.e., CTLHSO, can be well understood by the flowchart shown in Fig. 1.

The CTLHSO algorithm parameters are detailed in Table 3.

2.5 Pseudo-code of the CTLHSO algorithm

Input:	Read algorithm and distribution network data
Output:	Optimal strategy for simultaneous allocation of distributed generation and network reconfiguration
Step 1	Formulate the objective functions considering equality and inequality constraints
Step 2	Initialize the optimization parameters and independent variable constraints
Step 3	Randomly generate matrix of the initial solution vectors (SV) and determine its objective function values as formulated in *Step 1:*
Step 4	Set iteration count as 1
Step 5	Find auto-selection rate (ASR) as per Eq. (15)
Step 6	Select first SV from the initial solution vector matrix (SVM)
Step 7	Calculate objective function values corresponding to $X_{\text{new}}^{\text{SV}}$. Accept and update the SVM if $X_{\text{new}}^{\text{SV}}$ yields better objective function values than $X_{\text{old}}^{\text{SV}}$.
Step 8	*Learning phase* as per Eq. (4) is used for generation of the new SV $(X_{\text{new}}^{\text{SV}})$. Calculate objective function values corresponding to $X_{\text{new}}^{\text{SV}}$. Accept and update the SVM if $X_{\text{new}}^{\text{SV}}$ yields better objective function values than $X_{\text{old}}^{\text{SV}}$.
Step 9	Repeat 'Step 6:' to 'Step 8:' for each SV in SVM
Step 10	Update the HMCR^{ci}, PAR^{Mod} and DBW parameters in each iteration as per Eqs. (8), (10) and (13)
Step 11	Increase iteration count by 1
Step 12	Repeat 'Step 5:' to 'Step 10:' till the termination criterion of algorithm is met
Step 13	The SV '$X_{\text{new}}^{\text{SV}}$' having best objective function value is selected as solution of defined problem.

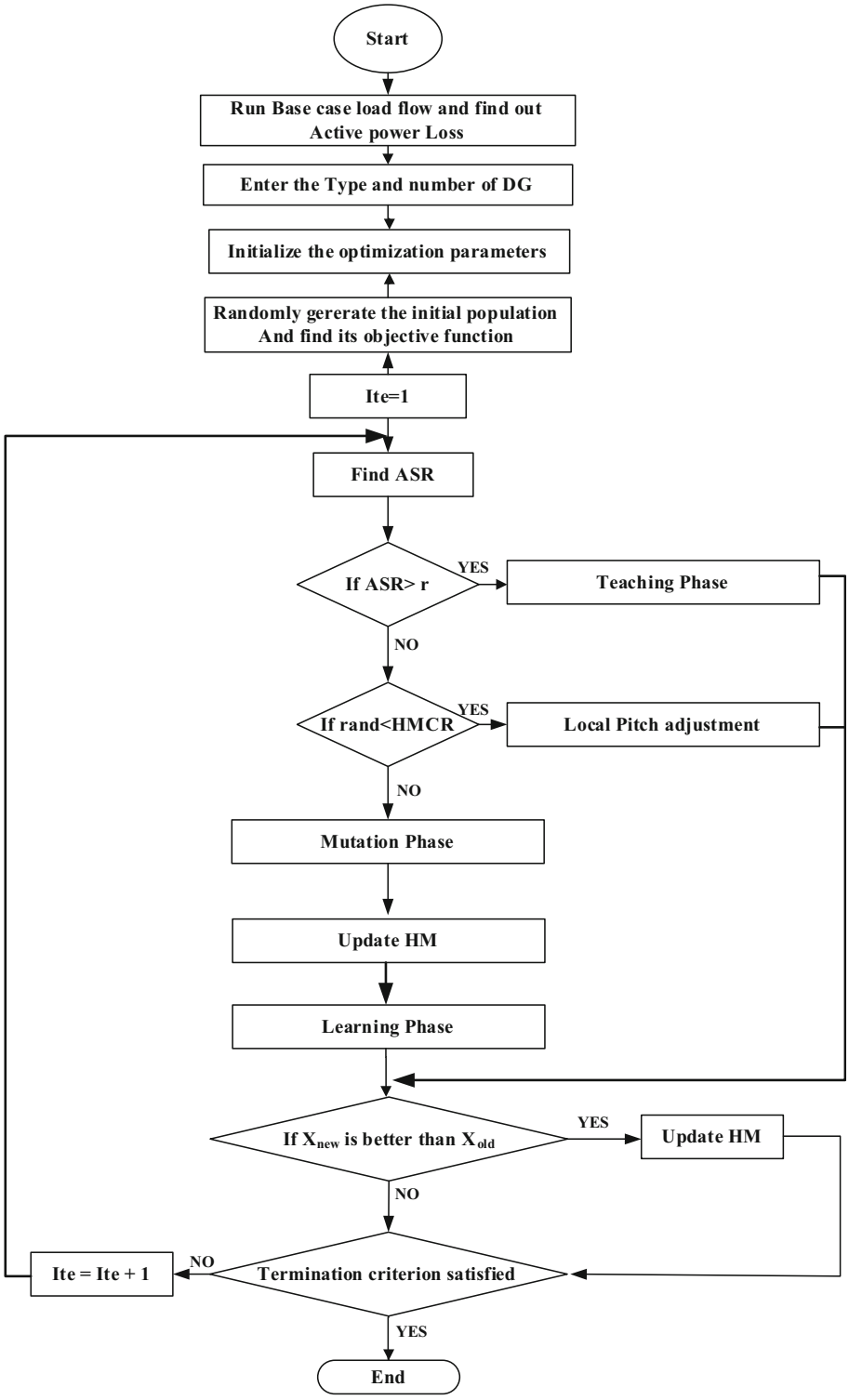


Fig. 1 Flowchart for the proposed CTLHSO algorithm

3 Problem formulation

The proposed technique is implemented on a deterministic problem of optimal DG sizing and placement in radial distribution systems addressing some core issues like active power losses, voltage deviation with and without network reconfiguration.

The proposed work has considered the following assumptions:

1 All RDSs are balanced.

2 Unity power factor of DGs has been employed in RDS.
3 Nominal load level and constant power load model are assumed.
4 The outputs of all DGs are time invariant.
5 The uncertainty of DER is not considered

3.1 Network power loss

Several load flow methods are available in the literature for the distribution network. This paper calculates active power loss of the distribution network using a direct load flow approach (Gözel and Hocaoglu 2009). This load flow approach is based on network topographical structure. The total active power loss of the distribution network is calculated using Eq. (16) and illustrated by Fig. 2a, as shown below:

$$P_{\text{loss}} = \sum_{j=1}^{\text{nb}} I_j^2 R_j \quad \text{and} \quad Q_{\text{loss}} = \sum_{j=1}^{\text{nb}} I_j^2 X_j$$
$$\text{where} \quad I_j = \sum_{k=1}^{m} \left(\left(H * P_{L,k}\right)^2 + H * Q_{L,k}^2 \right) / \left|V_k^2\right| \tag{16}$$

3.1.1 Power loss reduction after network reconfiguration

Network reconfiguration results in inclusion or exclusion of some tie/sectionalize switches. So after network reconfiguration, the BIBC and the BCBV matrices used in the direct load flow approach are modified. Subsequently, the total active power loss $\left(P_{\text{loss}}^{r}\right)$ is calculated using modified BIBC and BCBV using Eq. (16).

3.1.2 Active power loss after DG placement without reconfiguration

After placement of a DG, the active and reactive power demands at bus k in Eq. (16) are modified as follows:

$$P_{L,k}^{'} = P_{L,k} - P_{\text{DG},k} \quad \text{and} \quad Q_{L,k}^{'} = Q_{L,k} - Q_{\text{DG},k} \tag{17}$$

Thereafter, $P_{L,k}^{'}$ and $Q_{L,k}^{'}$ are substituted in place of $P_{L,k}$ and $Q_{L,k}$, respectively, in Eq. (16) to yield the power loss $(P_{\text{loss}}^{\text{DG}})$ of the network after DG placement.

3.1.3 Active power loss after DG placement and network reconfiguration

DG allocation in the distribution network causes change in the active and reactive power demands at bus k as detailed in Eq. (17). Subsequently due to network reconfiguration, BIBC and BCBV matrices are modified. Thereafter, $P_{L,k}^{'}$ and $Q_{Lk}^{'}$ in place of $P_{L,k}$ and $Q_{L,k}$, respectively, and modified BIBC and BCBV matrices are substituted in load flow, and active power loss of the network $(P_{\text{loss}}^{\text{DG},r})$ is calculated using Eq. (16).

3.2 Objective function

The objective of DG allocation and network reconfiguration in the distribution network is to minimize active power loss of the network while retaining all the constraints under different load levels. The minimization of active power loss of the distribution network is formulated in terms of fitness and is given by Eqs. (18) and (19).

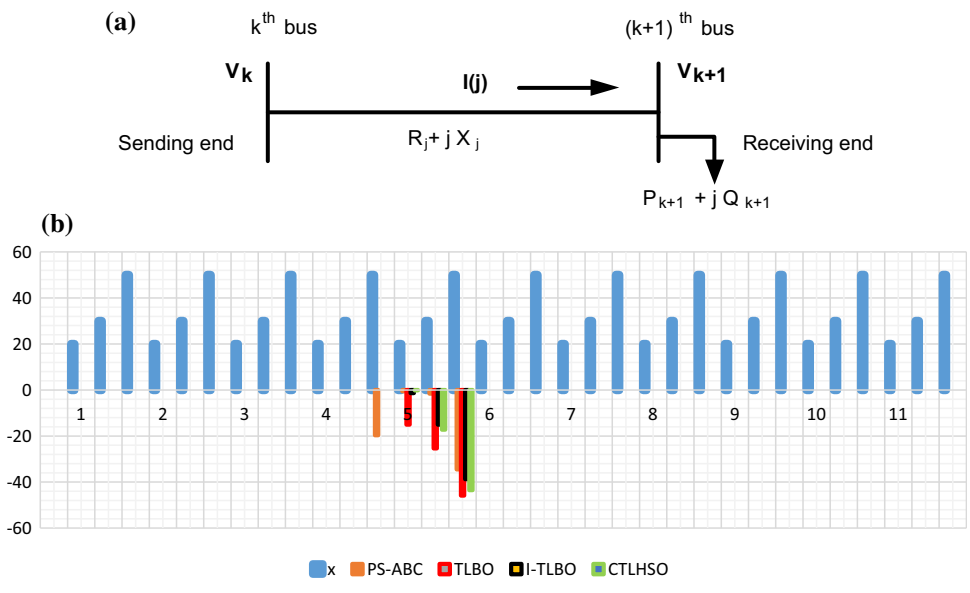


Fig. 2 **a** Equivalent circuit of the jth branch of the network between buses 'k' and '$(k+1)$'. **b** Comparative analysis of different optimization algorithms for 11 mathematical benchmark functions for 20, 30 and 50 variables

$$\text{Fitness} = P^{r}_{\text{loss}}/P^{\text{DG}}_{\text{loss}}/P^{\text{DG},r}_{\text{loss}} + \sum_{k=1}^{n}\left\{(|V_k| - V_{\min})^2 - (|V_k| - \text{V}_{\max})^2\right\} \quad (18)$$

$$\text{Objective function}(f) = \text{minimize (Fitness)} \quad (19)$$

Subject to the constraints given by Eqs. (20)–(25)

$$0.9\text{p.u.} \le V_k \le 1.05\text{p.u.and } I_j \le I_j^{\max} \quad (20)$$

$$\det(A) = 1 \text{ or } -1(\text{radial system}) \text{ or } \det(A) = 0(\text{not radial system}) \quad (21)$$

$$P^{\min}_{\text{DG},k} \le P_{\text{DG},k} \le P^{\max}_{\text{DG},k} \text{ where } 0 \le P_{\text{DG},k} \le 3\text{MW} \quad (22)$$

$$P_{\text{grid}} = \sum_{k=1}^{n} H * P_{L,k} + P^{r}_{\text{loss}} \quad (23)$$

$$\sum_{x=1}^{n\text{DG}} P_{\text{DG},x} + P_{\text{grid}} = \sum_{k=1}^{n} H * P_{L,k} + \left(P^{\text{DG}}_{\text{loss}}/P^{\text{DG},r}_{\text{loss}}\right) \quad (24)$$

$$Q_{\text{grid}} = \sum_{k=1}^{n} H * Q_{L,k} + (Q_{\text{loss}}/Q^{r}_{\text{loss}}/Q^{\text{DG}}_{\text{loss}}/Q^{\text{DG},r}_{\text{loss}}) \quad (25)$$

Reconfiguration and optimal placement of DGs in a distribution system result in the reduction in real power loss, voltage profile improvement and mitigation of network congestion. This improves the overall energy efficiency and lowers the operational cost of the distribution system with an added advantage of a reduction in the environmental impact

4 Realization of solution vector using CTLHSO algorithm for application in distribution networks

This section illustrates the application of CTLHSO algorithm for the network reconfiguration and DG allocation in the distribution networks under different scenarios to meet the objective function while satisfying all the constraints. Several attempts had been made for this complex optimization objective function by many researchers for both reconfiguration and optimal DG placement independently and simultaneously. In this paper, reconfiguration and multiple DG placements in the radial distribution system are taken simultaneously to minimize real power loss and improve voltage profile of the distribution networks using CTLHSO.

For the simultaneous reconfiguration and DG (only active power injection) placement, the number of variables in the solution vector (SV) will be equal to a number of tie switches (TS) plus two times number of DG, i.e., one variable for its location and one variable for its size. So after initialization of CTLHSO algorithm, SVM is generated according to Eq. (1) while satisfying all the constraints. This can be illustrated by considering an example of distribution network, i.e., IEEE 33-bus system, which has 5 tie switches and 2 DGs, which are considered to place in the network. So variables in solution vector will be 9. The solution vector of SVM can be represented as follows,

$$\text{SV} = \left[\begin{array}{ccccc|cc|cc} S^1_1 & S^1_2 & \overset{\longleftrightarrow}{S^1_3} & S^1_4 & S^1_5 & \text{DGL}^1_1 & \overset{\longleftrightarrow}{\text{DGL}^1_2} & \text{DGS}^1_1 & \overset{\longleftrightarrow}{\text{DGS}^1_2} \\ \multicolumn{5}{c|}{\text{Reconfiguration}} & \multicolumn{2}{c|}{\text{DG Location}} & \multicolumn{2}{c}{\text{DG size}} \end{array}\right] \quad (26)$$

where S^1_1, S^1_2, S^1_3, S^1_4, S^1_5 are the open tie or sectionalize switches in the distribution network corresponding to tie or sectionalize switches, provided that the radial nature of the network is maintained. DGL^1_1, DGL^1_2 are the locations of DGs, and DGS^1_1, DGS^1_2 are the size of DG in kW corresponding to their locations. Here the decision variables (X^{SV}_i) of solution vector (X^{SV}) representing tie or sectionalize switches and DG locations are discrete in nature, while the size of DG is a continuous number. In the same way, all the SV of SVM is generated and represented as shown in Eq. (2), while maintaining the radial structure of the network and satisfying other constraints. After initialization of SVM, in each iteration, CTLHSO algorithm is utilized to generate new solution vector using Eqs. (1), (4), (6), (14) and (15). If the fitness of new decision vector ($X^{\text{SV}}_{\text{new}}$) is better, update the SVM. This procedure of generation and updating of SVM is repeated until the termination condition is satisfied. The solution vector corresponding to the best fitness in SVM at the end of termination of optimization technique is presented as a solution of the objective function.

5 Results and performance evaluation

The proposed CTLHSO was primarily implemented on eleven standard mathematical benchmark functions for validation. The details of these functions and the results are shown in Tables 1 and 2, respectively. Comparative results in the form of mean value and standard deviation of eleven mathematical benchmark functions, when subjected to 120,000 maximum function evaluations for 30 independent runs, validate that the proposed optimization technique is superior or equivalent to Song (2013).

All considered cases are implemented using CTLHSO algorithm and programmed in MATLAB R2015a, and simulation is carried out on a computer on an Intel i5-4570, 3.2 GHz processor, 4 GB RAM, desktop PC.

Table 1 Mathematical benchmark functions (Song 2013)

S. no.	Function	Formulation	D	Search range
1.	Sphere	$F(x)_{\min} = \sum_{i=1}^{D} x_i^2$	30	[− 100, 100]
2.	Schwefel 2.22	$F(x) = \sum_{i=1}^{D} \lvert x_i \rvert + \prod_{i=1}^{n} \lvert x_i \rvert$	30	[− 10, 10]
3.	Schwefel 1.2	$F(x) = \sum_{i=1}^{D} \left(\sum_{j=1}^{i} x_j \right)^2$	30	[− 100, 100]
4.	Schwefel 2.21	$F(x) = \max_{1 \le i \le D} \lvert x_i \rvert$	30	[− 100,100]
5.	Rosenbrock	$F(x)_{\min} = \sum_{i=1}^{D} \left[100\left(x_i^2 - x_{i+1}\right)^2 + (1 - x_i)^2 \right]$	30	[− 2.048,2.048]
6.	Step 2	$F(x) = \sum_{i=1}^{D} \left([x_i + 0.5] \right)^2$	30	[− 100,100]
7.	Rastrigin	$F(x)_{\min} = \sum_{i=1}^{d} \left[x_i^2 - 10\cos(2\pi x_i) + 10 \right]$	30	[− 5.12,5.12]
8.	Ackley	$F(x)_{\min} = -20\mathrm{e}^{\left(-0.2\sqrt{\frac{1}{D}\sum_{i=1}^{D} x_i^2} \right)} - \mathrm{e}^{\left(\frac{1}{D}\sum_{i=1}^{D} \cos(2\pi x_i) \right)} + 20 + \mathrm{e}^1$	30	[− 32.768,32.768]
9.	Griewank	$F(x)_{\min} = \frac{1}{4000}\sum_{i=1}^{D} x_i^2 - \prod_{i=1}^{D} \cos\left(\frac{x_i}{\sqrt{i}} \right) + 1$	30	[− 600,600]
10.	Penalized	$F(x) = \frac{\pi}{D}\left[10\sin^2(\pi y_1) + \sum_{i=1}^{D-1} (y_i - 1)^2 \left\{ 1 + 10\sin^2(\pi y_{i+1}) \right\} + (y_D - 1)^2 \right] + \sum_{i=1}^{D} u(x_i, 10, 100, 4)$ $u(x_i, ak, m) = \begin{cases} k(x_i - a)^m, & x_i > a, \\ 0, -a \le x_i \le a, & y_i = 1 + 1/4(x_i + 1) \\ k(-x_i - a)^m & x_i < -a \end{cases}$	30	[− 50,50]
11.	Penalized 2	$F(x) = 0.1\Big[10\sin^2(\pi x_1) + \sum_{i=1}^{D-1} (x_i - 1)^2 \left\{ 1 + \sin^2(3\pi x_{i+1}) \right\} + (x_D - 1)^2$ $+ \left(1 + \sin^2(2\pi x_D) \right) \Big] + \sum_{i=1}^{D} u(x_i, 5, 100, 4)$ $u(x_i, ak, m) = \begin{cases} k(x_i - a)^m, & x_i > a, \\ 0, & -a \le x_i \le a \\ k(-x_i - a)^m & x_i < -a \end{cases}$	30	[− 50,50]

Table 2 Comparative results of CTLHSO algorithm with other algorithms (Song 2013) over 30 runs

F. no	x	G-ABC		PS-ABC		TLBO		I-TLBO		CTLHSO	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	20	3.1943E−16	7.3909E−17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	30	6.2643E−16	1.0859E−16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	50	1.2546E−5	6.0511E−9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	20	9.3611E−16	1.3278E−16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	30	1.3019E−10	4.6859E−11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	50	2.3671E−05	6.1889E−06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	20	2.6919E3	1.4619E3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	30	1.0939E4	2.5670E3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	50	4.1236E4	5.8269E3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	20	0.3325	1.0786	0.00	0.00	0.00	0.00	0.00	0.00	5.2307E−324	0.00
	30	12.6211	2.6556	8.59E−115	4.71E−114	4.9E−324	0.00	0.00	0.00	5.2307E−324	0.00
	50	45.3075	4.3151	19.6683	6.341E+00	9.9E−324	0.00	0.00	0.00	5.2307E−324	0.00
5	20	1.6769	2.9037	.5190	1.0764	15.0536	2.28E−01	1.3785	8.49E−01	**0.1412**	**0.14522**
	30	7.47961	19.0926	1.5922	4.4066	25.4036	3.50E−01	15.032	1.2E+00	17.232	2.3223
	50	25.7164	31.75811	34.4913	30.3412	45.8955	2.89E−01	38.7294	7.57E−01	43.522	2.3415
6	20	3.3386E−16	1.0154E−16	2.6147E−16	3.8684E−17	9.24E−33	4.36E−33	0.00	0.00	0.00	0.00
	30	6.4499E−16	1.1126E−16	5.7169E−16	8.2549E−17	1.94E−29	1.88E−29	0.00	0.00	3.1045−32	8.0539e32
	50	5.6529E−09	3.6854E−09	1.1674E−15	1.4114E−16	3.26E−13	5.11E−13	1.51E−32	8.89E−33	1.8917e−14	5.6245e−14
7	20	0.00	0.00	0.00	0.00	6.41E−14	6.16E−14	0.00	0.00	0.00	0.00
	30	3.3165E−02	1.8165E−01	0.00	0.00	6.95E−13	1.64E−12	0.00	0.00	0.00	0.00
	50	2.1733	1.0728	0.00	0.00	7.90E−13	1.89E−12	0.00	0.00	0.00	0.00
8	20	2.7533E−14	3.5832E−15	8.8817E−16	0.00	3.55E−15	8.32E−31	7.11E−16	0.00	3.4501e−15	0.00
	30	7.7828E−10	2.9817E−10	8.8817E−16	0.00	3.55E−15	8.32E−31	7.11E−16	0.00	3.4501e−15	0.00
	50	1.1137E−04	3.8873E−05	8.8817E−16	0.00	3.55E−15	8.32E−31	7.11E−16	0.00	3.4501e−15	0.00
9	20	6.0279E−04	2.2313E−03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	30	6.9655E−04	2.2609E−03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	50	1.0470E−03	2.7482E−03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	20	3.2621E−16	6.6721E−17	2.5576E−16	4.9715E−17	4.00E−08	6.85E−24	2.42E−16	1.09E−16	**3.4751e−32**	**6.2458e−48**
	30	5.8570E−16	1.1349E−16	5.5312E−16	8.6858E−17	2.67E−08	6.79E−12	4.98E−16	2.14E−16	**4.5245E−29**	**7.2541E−29**
	50	9.3017E−11	7.9664E−11	1.0252E−15	1.5815E−16	5.18E−05	1.92E−04	9.19E−16	5.38E−16	**4.6745E−16**	**3.5687E−16**
11	20	6.5528E−08	2.4413E−07	2.3456E−18	2.2088E−18	2.34E−08	6.85E−24	1.93E−18	1.12E−18	**2.5263E−32**	**0**
	30	2.1724E−07	5.6676E−07	6.0601E−18	5.6064E−18	2.37E−08	4.91E−10	5.92E−18	4.74E−18	**4.5872E−28**	**5.0245E−28**
	50	8.8776E−07	1.5324E−06	5.0541E−17	1.5350E−16	1.52E−03	5.29E−03	4.87E−17	4.26E−17	4.6258E−15	4.2354E−16

Bold letter highlights the improved result in comparison to other methods

Table 3 CTLHSO algorithm parameters

Parameter	Values
a	3
AC_{th}	0.02
$HMCR_{min}$	0.6
$HMCR_{max}$	0.9
PAR_{min}	0.2
PAR_{max}	0.4
r	0.95
BW_{min}	0.01
BW_{max}	0.1
Maxiteration	300
SVM	90

5.1 Performance evaluation considering mathematical benchmark functions

The performance of the proposed CTLHSO algorithm is validated on 11 mathematical benchmark test functions (Geem and Kim 2018) listed in Table 1. To access the performance of the proposed algorithm, two measures were used. Mean and standard deviation (SD) of the variables serve to evaluate the performance of the proposed algorithm,

$$\text{Mean}(\bar{x}) = \frac{\sum x}{n}$$

$$\text{SD} = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

where x and n are the value of solution variable and set of solution variables, respectively.

Table 4 Simulation result of different cases on 33-bus

Case 33-bus	Load level		
	Light	Nominal	Heavy
Case I			
SW_{open}	33, 34, 35, 36, 37	33, 34, 35, 36, 37	33, 34, 35, 36, 37
P_{Loss} (kW)	47.07	202.68	575.36
V_m (p.u.)	0.9583	0.91309	0.8528
Case II			
SW_{open}	7, 9, 14, 32, 37	7, 9, 14, 32, 37	7, 9, 14, 32, 37
P_{Loss} (kW)	33.27	139.55	380.45
V_m (p.u.)	0.9698	0.9378	0.8967
Case III			
SW_{open}	33, 34, 35, 36, 37	33, 34, 35, 36, 37	33, 34, 35, 36, 37
DG_{size} (MW)/(Bus)	0.3740/(14)	0.7540/(14)	1.2194/(14)
	0.5430/(24)	1.0995/(24)	1.7877/(24)
	0.5280/(30)	1.0714/(30)	1.7470/(30)
P_{Loss} (kW)	17.33	71.46	190.18
V_m (p.u.)	0.9844	0.9687	0.9494
Case IV			
SW_{open}	7, 9, 14, 32, 37	7, 9, 14, 32, 37	7, 9, 14, 32, 37
DG_{size} (MW)/(Bus)	0.4619/(8)	0.9317/(8)	1.5071/(8)
	0.5296/(24)	1.0682/(24)	1.7276/(24)
	0.4693/(30)	0.9504/(30)	1.5449/(30)
P_{Loss} (kW)	14.36	58.88	155.48
V_{min} (p.u.)	0.9872	0.9741	0.9580
Case V			
SW_{open}	11, 28, 31, 33, 34	11, 28, 31, 33, 34	11, 28, 31, 33, 34
DG_{size} (MW)/(Bus)	0.4761/(7)	0.9569/(7)	1.5408/(7)
	0.3750/(17)	0.7230/(17)	1.2110/(17)
	0.6328/(25)	1.2796/(25)	2.0758/(25)
P_{Loss} (kW)	12.41	50.72	133.42
V_m (p.u.)	0.9869	0.9734	0.9568

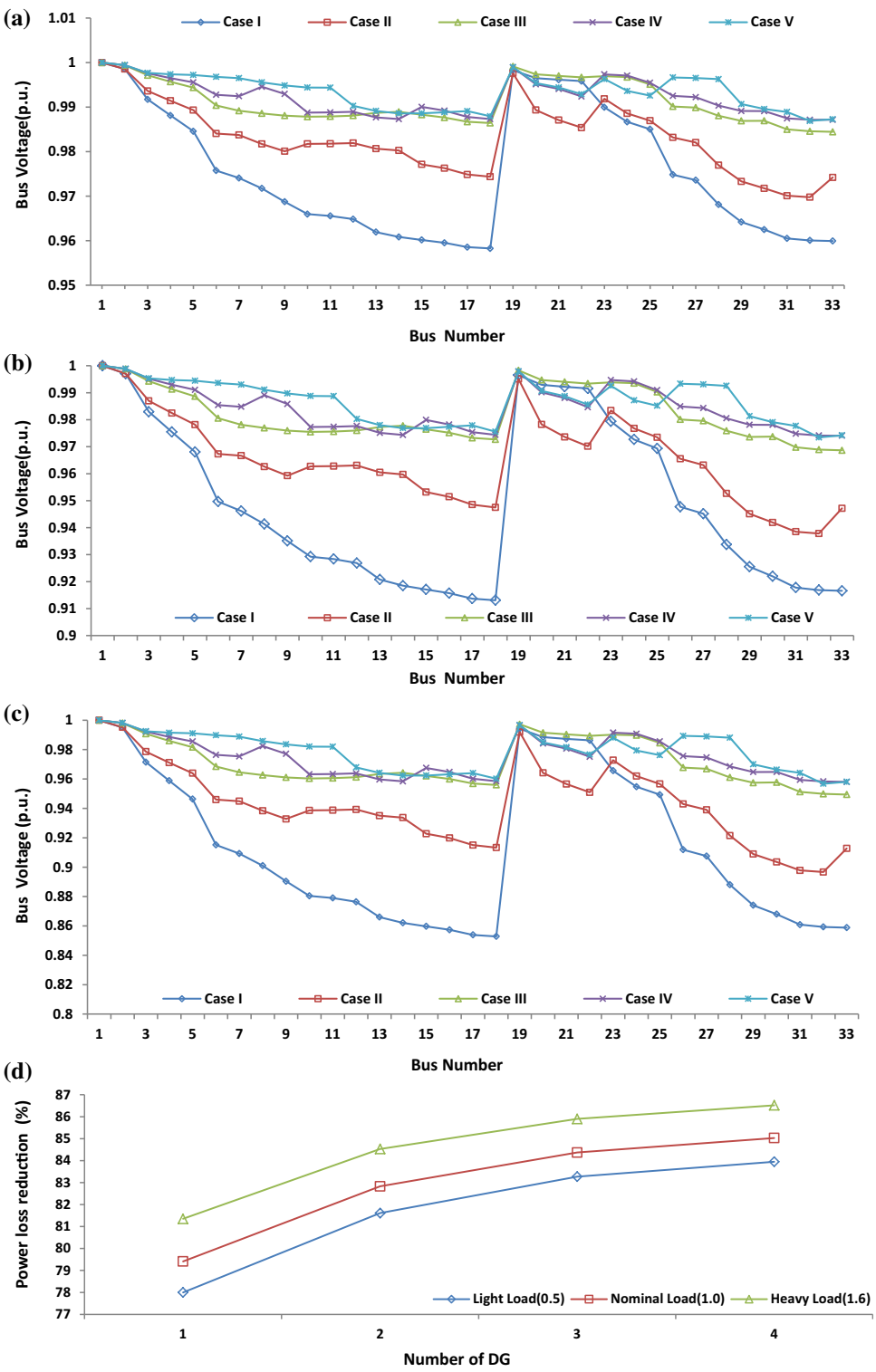
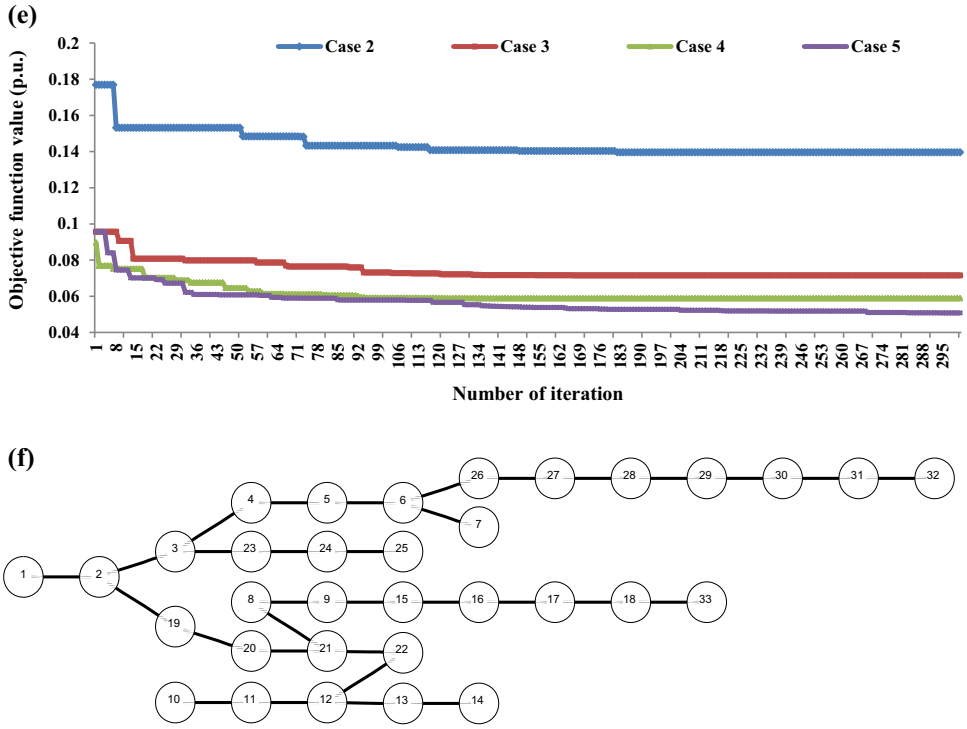


Fig. 3 **a** 33-Bus voltage profile for light load level (0.5). **b** 33-Bus voltage profile for nominal load level (1.0). **c** 33-Bus voltage profile for heavy load level (1.6). **d** Impact of number of DG on real power loss for 33-bus systems. **e** Convergence characteristics in different cases for 33-bus considering nominal load level (1.0). **f** Best reconfiguration for case 2 for 33-bus distribution network

A comparison of the proposed CTLHSO vis-à-vis other algorithms (Song 2013) is shown in Table 2. It is observed that the proposed method gives the best result with Penalized, Penalized 2 and Rosenbrock functions, while it yields results identical to I-TLBO (improved TLBO) for Sphere, Schwefel 2.22, Schwefel 1.2, Griewank, Ackley and Rastrigin, functions. However, it is observed that for the Schwefel 2.21, Step 2 function, I-TLBO gives the best result. It is observed from Fig. 2b that the proposed algorithm yields less error for the considered mathematical benchmark functions except Rosenbrock functions where CTLHSO has better result for 20 and 30 variables in

Fig. 3 continued



comparison with PS-ABC and TLBO but inferior to I-TLBO.

5.2 Algorithm application in distribution networks

The proposed CTLHSO is implemented on IEEE 33-bus and IEEE 69-bus test system, and its computational capability is demonstrated for different cases as follows,

Case I: Base case is taken where the network is without reconfiguration and DGs
Case II: The network system is reconfigured by proper switching of sectionalizing and tie switches, without DGs
Case III: The network with DGs placement without reconfiguration
Case IV: Optimal DGs placement after optimal network reconfiguration of the network
Case V: Simultaneous network reconfiguration and DGs placement

The developed algorithm, i.e., CTLHSO, is used for DG allocation and network reconfiguration for loss reduction and voltage profile improvement in the following distribution networks.

5.2.1 IEEE 33-bus distribution network

The data for IEEE 33-bus radial distribution system are taken from Baran and Wu (1989). The 33-bus system consists of 33 buses, 3 laterals, 37 branches with 5 loops or tie switches being kept generally open. The rated voltage is 12.66 kV with a total active and reactive power loadings of 3.72 MW and 2.3 MVAr, respectively. The distribution network is simulated for all the cases in three different load conditions, i.e., light load (0.5 p.u.), nominal load (1.0 p.u.), heavy load (1.6 p.u.), and the results are tabulated in Table 3. Table 4 briefly describes the impact of various cases on the radial distribution system for various load levels. Considering the base case, i.e., case I where distribution system is without DG and without reconfiguration, the real power losses (in kW) are 37.27, 139.55 and 575.36 for light, nominal and heavy loads, respectively. It is observed from Table 4 that for case IV (i.e., DG placement after reconfiguration), the open switch for reconfiguration is 7, 9, 14, 32, 37 and has real power losses (in kW) of 14.36, 58.88 and 155.48 for light, nominal and heavy loads, respectively, while case V (i.e., simultaneous reconfiguration and DGs placement) results in open switch 11, 28, 31, 33, 34 and real power losses (in kW) are 12.41, 50.72, 133.42 for light, nominal and heavy loads, respectively. This is the maximum reduction in real power loss, in comparison with all considered cases. The voltage profile of the IEEE 33-bus distribution system is shown in Fig. 3a–c for light, nominal and heavy loads considering all five cases. It is observed from the all the cases that in case V, voltage profile outcome is the best. Impact of number of DG for case V has been simulated for light, nominal and heavy loads. It is observed from Fig. 3d that real power loss of the network reduces nonlinearly with an

Table 5 Simulation result of different cases on 69-bus

Case 69-bus	Load level		
	Light	Nominal	Heavy
Case I			
SW_{open}	69, 70, 71, 72, 73	69, 70, 71, 72, 73	69, 70, 71, 72, 73
P_{Loss} (kW)	51.58	224.89	652.22
V_m (p.u.)	0.9567	0.9092	0.8445
Case II			
SW_{open}	14, 56, 61, 69, 70	14, 56, 61, 69, 70	14, 56, 61, 69, 70
P_{Loss} (kW)	23.60	98.57	267.02
V_m (p.u.)	0.9754	0.9495	0.9165
Case III			
SW_{open}	69, 70, 71, 72, 73	69, 70, 71, 72, 73	69, 70, 71, 72, 73
DG_{size} (MW)/(Bus)	0.2549/(11)	0.5268/(11)	0.8775/(11)
	0.1893/(18)	0.3796/(18)	0.6093/(18)
	0.8476/(61)	1.7190/(61)	2.7981/(61)
P_{Loss} (kW)	17.034	69.388	181.68
V_m (p.u.)	0.9894	0.9790	0.9669
Case IV			
SW_{open}	14, 56, 61, 69, 70	14, 56, 61, 69, 70	14, 56, 61, 69, 70
DG_{size} (MW)/(Bus)	0.2682/(11)	0.5376/(11)	0.8622/(11)
	0.7125/(61)	1.4415/(61)	2.3125/(61)
	0.2440/(64)	0.4900/(64)	0.7882/(64)
P_{Loss} (kW)	8.632	35.145	91.963
V_m (p.u.)	0.9907	0.9813	0.9699
Case V			
SW_{open}	14, 56, 61, 69, 70	14, 56, 61, 69, 70	14, 56, 61, 69, 70
DG_{size} (MW)/(Bus)	0.2682/(11)	0.5376/(11)	0.8622/(11)
	0.7125/(61)	1.4415/(61)	2.3125/(61)
	0.2440/(64)	0.4900/(64)	0.7882/(64)
P_{Loss} (kW)	8.632	35.145	91.963
V_m (p.u.)	0.9907	0.9813	0.9699

increase in the number of DG. So up to three DG placements for the case V in the 33-bus system will be economically feasible. Figure 3e shows the convergence characteristics for different cases. It is observed that case 3 and case 4 have faster convergence rate in comparison with case 2 and case 5. Figure 3f represents the reconfigured network which shows minimum active power loss in the 33-bus distribution network for case II.

5.2.2 IEEE 69-bus test system

The data for the IEEE 69-bus radial distribution system are taken from Das et al. (1994). The system consists of 69 buses, 7 laterals and 73 branches with five loops or tie switches being kept generally open. The rated voltage is 12.66 kV with a total active and reactive power loadings of 3.8 MW and 2.69 MVAr, respectively. This test case is simulated with the proposed CTLHSO with same parameter and similar three load conditions as in test case A presented in Table 5. It is observed that for the case I (i.e., base case), real power losses for this system are 51.58, 224.89 and 652.22 for light, nominal and heavy load levels, respectively. The study shows that case II suggests four switch open conditions, i.e., {14, 55, 61, 69, 70},{14, 56, 61, 69, 70}, {14, 57, 61, 69, 70} and {14, 58, 61, 69, 70}, for network reconfiguration and results in the same real power loss (in kW) of 23.6, 98.57 and 267.02 of the network for light (0.5), nominal (1.0) and heavy (1.6) load levels, respectively. It is also observed that results for case V are same as case IV.

The real power loss (in kW) for case V (i.e., with reconfiguration) when switches {14, 56, 61, 69, 70} are open and DGs are installed is 8.632, 35.145 and 91.963 for light (0.5), nominal (1.0) and heavy (1.6) load levels, respectively. The voltage profile of the IEEE 69-bus distribution system is shown in Fig. 4a–c for light, nominal

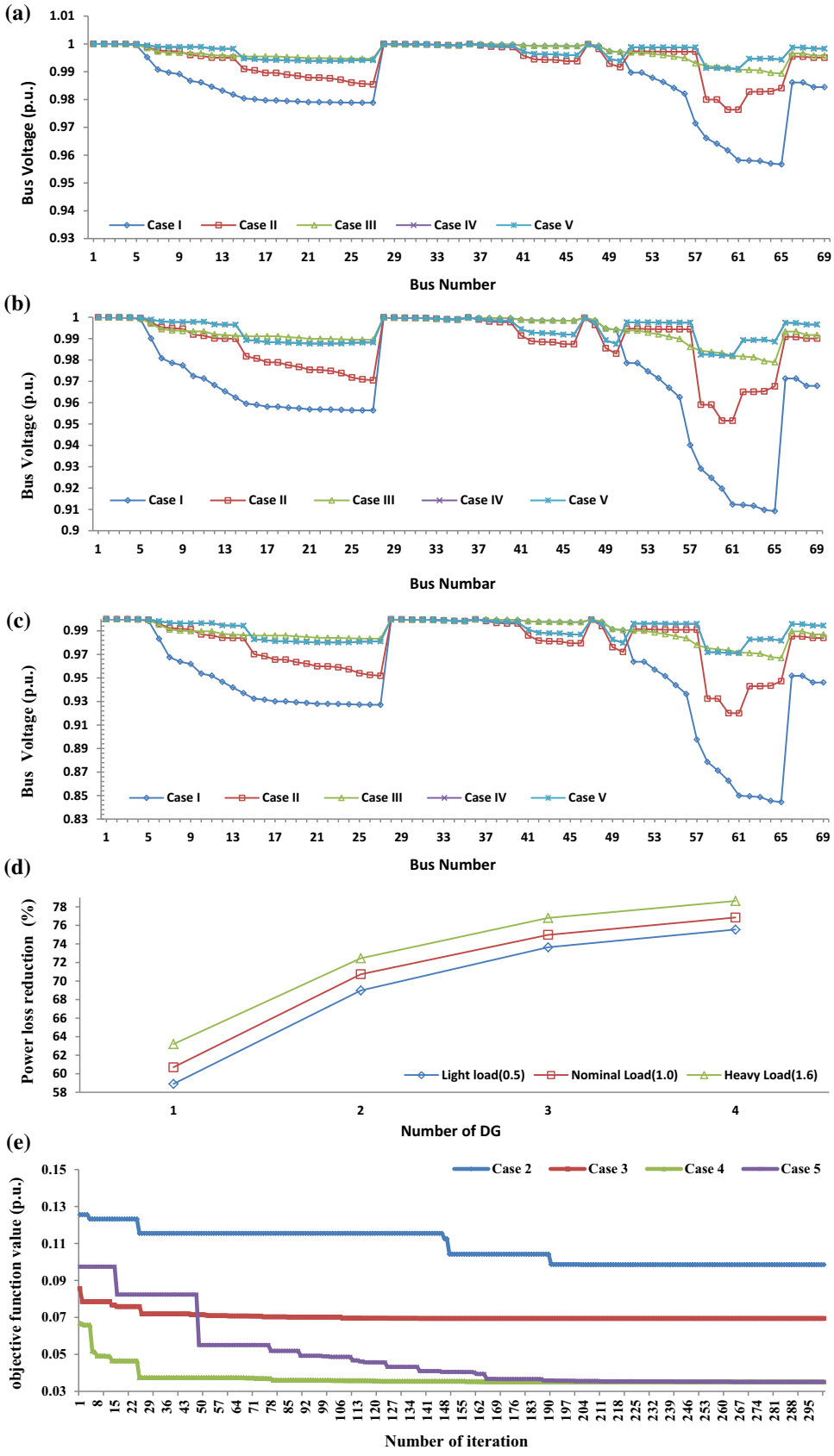


Fig. 4 **a** 69-Bus voltage profile for light load level (0.5). **b** 69-Bus voltage profile for nominal load level (1.0). **c** 69-Bus voltage profiles for heavy load level (1.6). **d** Impact of number of DG on real power loss for 69-bus. **e** Convergence characteristics in different cases for 69-bus considering nominal load level (1.0). **f** Best reconfiguration for case 2 for 69-bus distribution network

Fig. 4 continued

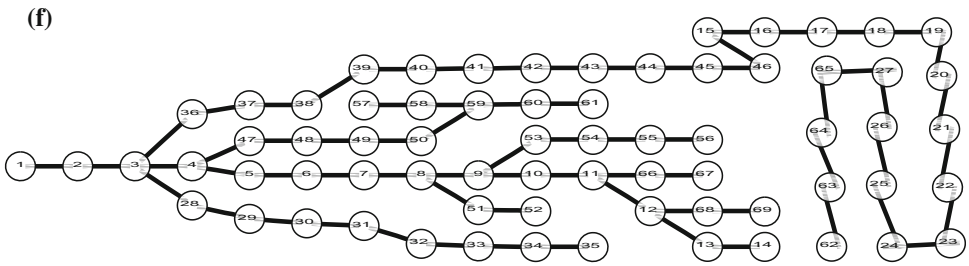


Table 6 Comparative study of simulation result for different cases on 33-bus

Algorithm	Description	Case II	Case III	Case IV	Case V
CTLHSO	SW_{open}	**7, 9, 14, 32, 37**	**33, 34, 35, 36, 37**	**7, 9, 14, 32, 37**	**7, 10, 14, 28, 32**
	P_{Loss} (kW)	**139.55**	**71.457**	**58.877**	**50.718**
	V_{min} (p.u.)	**0.9378**	**0.9687**	**0.9741**	**0.9734**
	DG_{size} (MW)	–	**2.9249**	**2.9502**	**2.9895**
FWA (Imran et al. 2014)	SW_{open}	7, 9, 14, 28, 32	33, 34, 35, 36, 37	7, 9, 14, 28, 32	7, 11, 14, 28, 32
	P_{Loss} (kW)	139.98	88.68	83.91	67.11
	V_{min} (p.u.)	0.9413	0.9680	0.9612	0.9713
	DG_{size} (MW)	–	1.7938	1.0728	1.6841
HSA (Srinivasa Rao et al. 2013)	SW_{open}	7, 9, 14, 32, 37	33, 34, 35, 36, 37	7, 9, 14, 32, 37	7, 10, 14, 28, 32
	P_{Loss} (kW)	138.06	96.76	97.13	73.05
	V_{min} (p.u.)	0.9342	0.9670	0.9479	0.9700
	DG_{size} (MW)	–	1.7256	1.0909	1.6684
GA (Srinivasa Rao et al. 2013)	SW_{open}	9, 28, 33, 34, 36	33, 34, 35, 36, 37	9, 28, 33, 34, 36	7, 10, 28, 32, 34
	P_{Loss} (kW)	141.60	100.1	98.36	75.13
	V_{min} (p.u.)	0.9310	0.9605	0.9506	0.9677
	DG_{size} (MW)	–	1.6044	1.448	1.9633
RGA (Srinivasa Rao et al. 2013)	SW_{open}	7, 9, 14, 32, 37	33, 34, 35, 36, 37	7, 9, 14, 32, 37	7, 9, 12, 27, 32
	P_{Loss} (kW)	139.46	97.60	98.23	74.32
	V_{min} (p.u.)	0.9315	0.9687	0.9479	0.9691
	DG_{size} (MW)	–	1.777	1.100	1.774

and heavy loads considering all five cases. It is observed from the all the cases that case V voltage profile outcome is best among all.

Impact of number of DGs for case V has been simulated for light, nominal and heavy loads. The increase in the number of DGs is observed from Fig. 4d. So up to three DG placements for the case V in the 69-bus system will be economically feasible. Figure 4e shows the convergence characteristics for different cases. It is observed that case 3 and case 4 have faster convergence rate in comparison with case 2 and case 5. Figure 4f represents the reconfigured network which shows minimum active power loss in the 69-bus DN for case II.

5.3 Comparing CTLHSO algorithm with other algorithm in distribution networks

The simulated result of DG allocation and network reconfiguration using the proposed CTLHSO algorithm in IEEE 33-bus and 69-bus for case V considering nominal load level (1.0) is presented in Tables 6 and 7, respectively, in comparison with HSA, GA, RGA (Srinivasa Rao et al. 2013) and FWA (Imran et al. 2014). It is observed from Tables 6 and 7 that CTLHSO algorithm yields maximum power loss reduction in case V, i.e., 63.5% and 64.34% for the 33-bus and 69-bus distribution network, respectively. Further, it is observed that proper siting and sizing along with the reconfiguration of networks result in less voltage deviation in all the cases. Figure 5a, b shows that CTLHSO algorithm gives better performance for case III, case IV and case V than FWA and HSA algorithms in the context

Table 7 Comparative study of simulation result for different cases on 69-bus

Algorithm	Description	Case II	Case III	Case IV	Case V
CTLHSO	SW_{open}	**14, 56, 61, 69, 70**	**69, 70, 71, 72, 73**	**14, 56, 61, 69, 70**	**14, 56, 61, 69, 70**
	P_{Loss} (kW)	**98.572**	**69.388**	**35.145**	**35.145**
	V_{min} (p.u.)	**0.9495**	**0.9790**	**0.9813**	**0.9813**
	DG_{size} (MW)	–	**2.6253**	**2.4616**	**2.4616**
FWA (Imran et al. 2014)	SW_{open}	14, 56, 61, 69, 70	69, 70, 71, 72, 73	14, 56, 61, 69, 70	13, 55, 63, 69, 70
	P_{Loss} (kW)	98.59	77.85	43.88	39.25
	V_{min} (p.u.)	0.9495	0.9740	0.9720	0.9796
	DG_{size} (MW)	–	1.8329	1.3584	1.8182
HSA (Srinivasa Rao et al. 2013)	SW_{open}	13, 18, 56, 61, 69	69, 70, 71, 72, 73	13, 18, 56, 61, 69	13, 17, 58, 61, 69
	P_{Loss} (kW)	99.35	86.77	51.30	40.30
	V_{min} (p.u.)	0.9428	0.9677	0.9619	0.9736
	DG_{size} (MW)	–	1.7732	1.8448	1.8718
GA (Srinivasa Rao et al. 2013)	SW_{open}	14, 53, 61, 69, 70	69, 70, 71, 72, 73	14, 53, 61, 69, 70	10, 15, 45, 55, 62
	P_{Loss} (kW)	103.29	88.5	54.53	46.20
	V_{min} (p.u.)	0.9411	0.9687	0.9401	0.9727
	DG_{size} (MW)	–	1.9471	1.7422	2.0292
RGA (Srinivasa Rao et al. 2013)	SW_{open}	13, 17, 55, 61, 69	69, 70, 71, 72, 73	13, 17, 55, 61, 69	10, 14, 16, 55, 62
	P_{Loss} (kW)	100.28	87.65	52.34	44.23
	V_{min} (p.u.)	0.9428	0.9678	0.9611	0.9742
	DG_{size} (MW)	–	1.7868	1.6396	2.0654

of real power loss. The convergence characteristics of the CTLHSO algorithm in comparison with HSA and GA are shown in Fig. 5c. The proposed algorithm results for all the cases in Tables 6 and 7 are in bold.

5.4 Merits and limitations of CTLHSO algorithm

The CTLHSO algorithm having merits of both meta-heuristic algorithms, i.e., good exploration capability along with faster convergence characteristics of TLBO and better exploitation characteristics of the HS algorithm, yields better results for DGs allocation and network reconfiguration in distribution networks. The proposed algorithm has shown better result while handling mix integer variables (i.e., DGs location and reconfiguration are discrete variables, while DGs size is continuous variables) and constraints (i.e., voltage and DG size constraints), in comparison with HSA, RGA, FWA and GA.

Although TLBO algorithm is almost parameter independent, HS algorithm needs to be properly tuned for global optimization and faster convergence. Hence, the proposed algorithm also required to be properly tuned for the global solution. So, after several runs, the best-tuned algorithm parameter values have been considered as shown in Table 3 for the proposed algorithm to allocate DGs and reconfiguration of the distribution networks.

6 Conclusions and scope for further work

This paper proposes a new population-based optimization algorithm for simultaneous network reconfiguration and optimal DG placement to minimize real power losses in radial distribution systems. As compared to TLBO and HSA, the proposed technique is faster and immune to local extrema trappings. Initially, the proposed algorithm, i.e., CTLHSO, is validated using several mathematical benchmark functions. Subsequently, several case studies were carried out vis-à-vis combinations of network reconfiguration and optimal placement of DGs in the IEEE 33-bus and 69-bus radial distribution test systems. These include cases with exclusive DG placements and/or network reconfigurations, DG placement after reconfiguration and simultaneous DG placement with network reconfiguration. Three different load levels were considered. Results with the proposed CTLHSO algorithm show higher power loss reduction and improved network voltage profiles for both IEEE 33-bus and 69-bus distribution systems, in comparison with FWA, GA, RGA and HSA algorithms. It is also observed that (a) simultaneous DG placement and network reconfiguration yield better results and (b) real power losses in distribution systems decrease with increase in the number of DGs up to a certain limit, beyond which the improvement is rendered marginal.

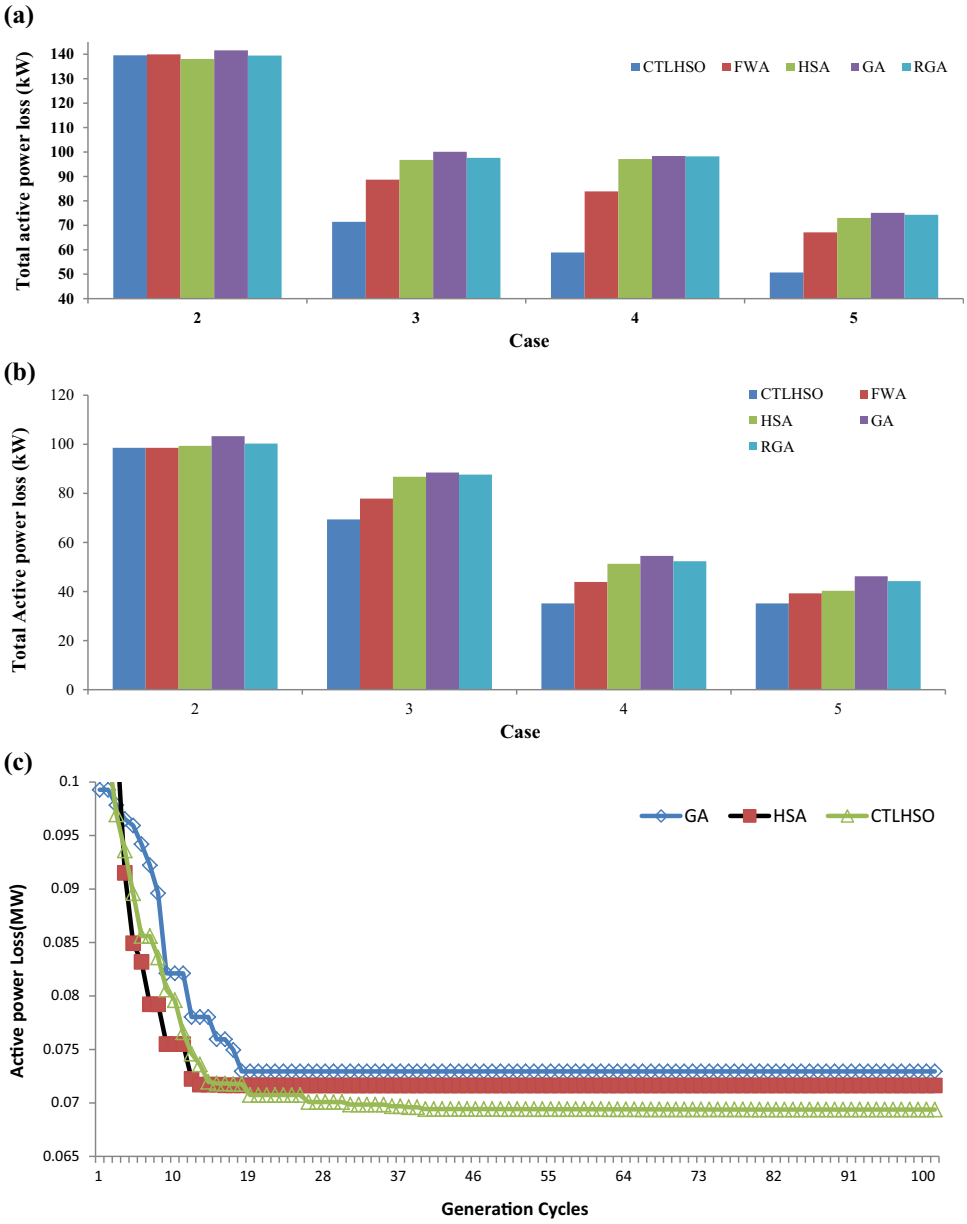


Fig. 5 **a** Comparison of active power loss in different cases for 33-bus. **b** Comparison of real power loss in different cases for 69-bus. **c** Comparison of convergence characteristics in case 3 for 69-bus considering nominal load level

Although in this paper the proposed algorithm has considered single-objective problem with constraints, it can be utilized for multi-objective function evaluation as well as distribution network planning where a large number of variables are involved, along with several constraints.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval The article does not contain any studies with human participants or animals performed by any of the authors.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

Abu-Mouti FS, El-Hawary ME (2011) Optimal distributed generation allocation and sizing in distribution systems via artificial bee colony algorithm. IEEE Trans Power Deliv 26(4):2090–2101

Ali SH (2012) A novel tool (FP-KC) for handle the three main dimensions reduction and association rule mining. In: IEEE, 2012 6th international conference on sciences of electronics, technologies of information and telecommunications (SETIT), Sousse, pp 951–961

Al-Janabi S (2018) Smart system to create optimal higher education environment using IDA and IOTs. Int J Comput Appl **(in press)**

Al-Janabi S, Alkaim AF (2019) A nifty collaborative analysis to predicting a novel tool (DRFLLS) for missing values estimation. Soft Comput. https://doi.org/10.1007/s00500-019-03972-x

Al-Janabi S, Ibrahim A-S (2017) Assessing the suitability of soft computing approaches for forest fires prediction. Appl Comput Inform. https://doi.org/10.1016/j.aci.2017.09.006

Al-Janabi S, Sarvesh R, Ahmed P, Ibrahim A-S (2015) Design and evaluation of a hybrid system for detection and prediction of faults in electrical transformers. Int J Electr Power Energy Syst 67:324–335

Al-Janabi S, Salman MA, Fanfakh A (2018) Recommendation system to improve time management for people in education environments. J Eng Appl Sciences 13(24):10182–10193

Ameli A, Bahrami S, Khazaeli F, Haghifam MR (2014) A multiobjective particle swarm optimization for sizing and placement of DGs from DG owners' and distribution company's viewpoints. IEEE Trans Power Deliv 29(4):1831–1840

Baran ME, Wu FF (1989) Network reconfiguration in distribution systems for loss reduction and load balancing. IEEE Trans Power Deliv 4(2):1401–1407

Chen J, Pan Q, Li J (2012) Harmony search algorithm with dynamic control parameters. Appl Math Comput 219(2):592–604

Das D, Nagi HS, Kothari DP (1994) Novel method for solving radial distribution networks. IEE Proc Gener Transm Distrib 141(4):291–298

Das S, Mukhopadhyay A, Roy A, Abraham A, Panigrahi BK (2011) Exploratory power of the harmony search algorithm: analysis and improvements for global numerical optimization. IEEE Trans Syst Man Cybern B 41(1):89–106

Geem ZW, Kim J-H (2018) Application of computational intelligence techniques to an environmental flow formula. Int J Fuzzy Log Intell Syst 18(4):237–244

Geem ZW, Kim JH, Loganathan GV (2001) A new heuristic optimization algorithm: harmony search. Simulation 76(2):60–68

Gomez-Gonzalez M, López A, Jurado F (2012) Optimization of distributed generation systems using a new discrete PSO and OPF. Electr Power Syst Res 84(1):174–180

Gopiya Naik SN, Khatod DK, Sharma MP (2015) Analytical approach for optimal siting and sizing of distributed generation in radial distribution networks. IET Gener Transm Distrib 9(3):209–220

Gözel T, Hocaoglu MH (2009) An analytical method for the sizing and siting of distributed generators in radial systems. Electr Power Syst Res 79(6):912–918

Hung DQ, Mithulananthan N, Bansal RC (2010) Analytical expressions for DG allocation in primary distribution networks. IEEE Trans Energy Convers 25(3):814–820

Imran AM, Kowsalya M, Kothari DP (2014) A novel integration technique for optimal network reconfiguration and distributed generation placement in power distribution networks. Int J Electr Power Energy Syst 63:461–472

Kalajdzic K, Ali SH, Patel A (2015) Rapid lossless compression of short text messages. Comput Stand Interfaces 37:53–59

Kanwar N, Gupta N, Niazi KR, Swarnkar A (2015) Simultaneous allocation of distributed resources using improved teaching learning based optimization. Energy Convers Manag 103:387–400

Kattan A, Abdullah R (2013) A dynamic self-adaptive harmony search algorithm for continuous optimization problems. Appl Math Comput 219(16):8542–8567

Mahdavi M, Fesanghary M, Damangir E (2007) An improved harmony search algorithm for solving optimization problems. Appl Math Comput 188(2):1567–1579

Martín García JA, Gil Mena AJ (2013) Optimal distributed generation location and size using a modified teaching–learning based optimization algorithm. Int J Electr Power Energy Syst 50:65–75

Mohanty B, Tripathy S (2016) A teaching learning based optimization technique for optimal location and size of DG in distribution network. J Electr Syst Inf Technol 3(1):33–44

Moradi MH, Abedini M (2012) A combination of genetic algorithm and particle swarm optimization for optimal DG location and sizing in distribution systems. Int J Electr Power Energy Syst 34(1):66–74

Niknam T, Taheri SI, Aghaei J, Tabatabaei S, Nayeripour M (2011) A modified honey bee mating optimization algorithm for multiobjective placement of renewable energy resources. Appl Energy 88(12):4817–4830

Pan Q, Suganthan PN, Tasgetiren MF, Liang JJ (2010) A self-adaptive global best harmony search algorithm for continuous optimization problems. Appl Math Comput 216(3):830–848

Rao RV, Savsani VJ, Vakharia DP (2011) Computer-aided design teaching–learning-based optimization : a novel method for constrained mechanical design optimization problems. Comput Aided Des 43(3):303–315

Saha S, Mukherjee V (2016) Optimal placement and sizing of DGs in RDS using chaos embedded SOS algorithm. IET Gener Transm Distrib 10(14):3671–3680

Sheng W, Liu K, Liu Y, Ye X, He K (2016) Reactive power coordinated optimisation method with renewable distributed generation based on improved harmony search. IET Gener Transm Distrib 10(13):3152–3162

Singh D, Singh D, Verma KS (2008) GA based energy loss minimization approach for optimal sizing & placement of distributed generation. Int J Knowl-Based Intell Eng Syst 12(2):147–156

Song H (2013) Fuzzy-enforced complementarity constraints in nonlinear interior point method-based optimization. Int J Fuzzy Log Intell Syst 13:171–177

Srinivasa Rao R, Ravindra K, Satisk K, Narasimham SVL (2013) Power loss minimization in distribution system using network reconfiguration in the presence of distributed generation. IEEE Trans Power Syst 28(1):317–325

Sultana S, Roy PK (2014) Multi-objective quasi-oppositional teaching learning based optimization for optimal location of distributed generator in radial distribution systems. Int J Electr Power Energy Syst 63:534–545

Tuo S, Zhang J, Yong L, Yuan X, Liu B, Xu X (2015) A harmony search algorithm for high-dimensional multimodal optimization problems. Digit Signal Process 46:151–163

Wang C, Nehrir MH (2004) Analytical approaches for optimal placement of distributed generation sources in power systems. IEEE Trans Power Syst 19(4):2068–2076

Yadav P, Kumar R, Panda SK, Chang CS (2012) An intelligent tuned harmony search algorithm for optimisation. Inf Sci 196:47–72

Zou D, Gao L, Wu J, Li S (2010) Neurocomputing Novel global harmony search algorithm for unconstrained problems. Neurocomputing 73(16–18):3308–3318

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

A New Method for Semantic Similarity Assessment using Fuzzy Formal Concept Analysis & Fuzzy Set Similarity Measure

Shivani Jain, Seeja.K.R, Rajni Jindal

Abstract*: Measuring the accurate semantic similarity between the words is a major issue in various applications of artificial intelligence and computational linguistics areas such as natural language processing, text-mining, information retrieval and for development of semantic web. In the past, many approaches have been proposed and adopted to evaluate similarity by using the knowledge based systems such as WordNet and MeSH ontology. In this paper we have proposed a new method; based on hybridization approach in knowledge based system. In this we have used feature based method and fuzzy Set theory. In feature based approach, properties or features are used for measuring the similarity as compare to edge and content information approaches. Our approach is investigated on standard dataset like R&G, M&C and 353-TC, which shows prominent improvement in the judgment of semantic similarity score between the words. This approach can be further used among cross ontology and fuzzy ontology as it is based on the feature based measure and fuzzy set theory.

Index Terms: Semantic Similarity Measures, Formal concept Analysis, Fuzzy Formal Concept Analysis, Word Net

I. INTRODUCTION

Research is drowning interest towards the semantic web[1]. The semantic web has a shared understanding, a defined structural and extension of current World Wide Web[2]. Ontology play's an extremely important role for the development of a semantic web. Ontology is defined by the Author Gurber et al [3] "it is a formal specification of a shared conceptualization". All-though ontologies are created for the different aspects and domains, they regularly contain overlapping data and the information. This information can be further used to evaluate the similarity among the words[4] .Semantic similarity; measures a numerical value that specify the closeness among the words/concepts. The idea of similarity or likeness is to identify the concepts having common characteristics. Appropriate assessment of similarity improves the understanding of textual data[5].More semantically close words are those that share same idea or terms, specified by the synonym in the English language like "pain and hurt", two different words but can be used inter-changeably. To measure the accurate similarity among the words/concepts is the open research problem in the area of computational linguistic[6] and natural language processing[7]. In the fig1 Car, ship and bus which are different class, having 'is-a' relation with Automobile, in a travel ontology. Similarity can be computed between the {automobile, car},{automobile, bus},{bus, car} and so on..Semantic closeness shows how taxonomically near two Terms are, as they share some common attribute of their meaning. Two terms highly associated with the concept of semantic similarity is semantic distance and semantic relatedness.

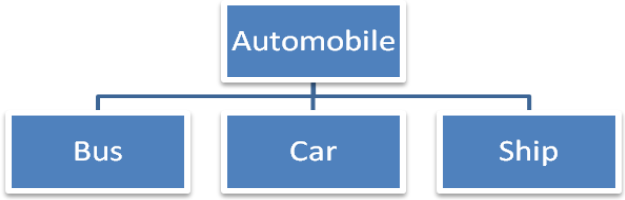


Figure 1: Showing "is-a" Relationship on Travel Ontology

Semantic distance [8] is computed as edge or link difference between the two terms, how far the two terms; like dist(bus, car)=dist(car, ship) but sim(bus, car) ≠ sim(car, ship). Another term associated with semantic similarity is 'semantic relatedness', which does not only rely on the taxonomic relation "is-a" more relations to be considered like part –of, antonym relationship's that are present in the WordNet [9] can be considered. Words {ink, pencil} are less related to each other as compared to {pencil, paper} in terms of semantic relatedness.

In this paper, proposed a new method for feature based approaches. In this WordNet is used for taxonomical information, Formal Concept Analysis (FCA) is exploit for extraction of features for given concept or words. Linguistics value to each feature is assigned to show how much an attribute/feature is related to the corresponding concept/word/class accordance with fuzzy logic and mapped into fuzzy formal concept Analysis. A multi-attribute features table is constructed for the concepts. A fuzzy set similarity measure is used to compute similarity among the words. The overall process yields us a better result in terms of accuracy on benchmark datasets. The same approach can be applied in Global ontologies, cross ontologies and fuzzy ontology.

The paper is organized in following sections. Section 2 describes the related work in the area of semantic similarity. Section 3 describes the proposed approach and different preliminary for the proposed approach.

Revised Version Manuscript Received on 30 November, 2018.

Shivani Jain, Department of Computer Science & Engineering, Indira Gandhi Delhi Technological University for women, Delhi, India

Seeja.K.R, Department of Computer Science & Engineering, Indira Gandhi Delhi Technological University for women, Delhi, India

Rajni Jindal, Department of Computer Science & Engineering, Delhi Technological University, Delhi, India

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Section 5, 6 presents the step followed in our approach and evaluation respectively. Result section shows the result on TC-353 similarity data-set and comparison with other methods. Last division illustrates the conclusion and future work.

II. RELATED WORK

In general two words can be similar lexically or semantically. Lexical similarity represents, words having similar string sequence like in {"Man", "Lan" and "Van"} and also in DNA sequence matching {"ADCGTDCGTC" and "ADCCGTCGCA"}. Lexical similarity has a wide application in the areas of medical sequence matching [10] and pattern recognition[11]. Various measures are proposed in this area[12]. Whereas semantic similarity deals with the meaning of the two words, how the linguistic meanings of two words are similar such as how much similar the words {"mango" and "orange"} as both words belongs to class "Fruit". For calculating the semantic similarity among the words diverse similarity measures have been proposed by different researchers groups. In literature two common approaches "Corpus-Based Approaches" and "Knowledge-Based Approaches" are presented by the researchers [13]. In corpus based approaches large amount of text context is used for calculating the similarity among the words. Latent Semantic Approach (LSA) [14] and Pointwise mutual information(PMI)[15] are the most used methods for computing the semantic similarity.

The major related work on semantic similarity measures utilizing the taxonomy knowledge by the Global ontologies like WordNet, MeSH[16] a medical oncology and SUMO ontology. Knowledge-based Methods [17] are an effort to figure out the semantic similarity without human intervention, it uses a vast amount of environment knowledge about the concepts. Ontologies have been extensively used in the area of knowledge-base systems. WordNet is a domain free and all-purpose thesaurus of common English words. It formulates approximately 10000 English words in a semantic structure that looks like ontology. The resultant of WordNet concepts forms a network of significantly related words. A graph prototype is generated in form of network and used for forming the meaning of concepts. WordNet is exploited as background ontology in largely associated areas of the knowledge-based system. Broadly classified Knowledge-based technique has three approaches [5], [13][18]and the last one is hybrid approaches which can be combination of any of these method. The approaches are (1)Edge/path based Approaches (2) Information Content approaches (3) Features Based Approaches (4)Hybrid Approaches. Edge-Counting approach introduced by Author et al Lin [19] after that many measures was introduced by the different researchers[20], [21],[22] it is a straight-forward technique to measure the similarity, it calculates the minimum path-distance from end to end connection with their related ontological models through a "is-a" link. As the path length increases the distance is also increased and less similar the word are. The drawback of these measures are that; they only work on the "is-a" link and works in the single ontology structure. In information Content Approach, it is based on the theory that every concept has a large extent of information in WordNet. In a multidimensional space where a node represents a unique concept it contains definite amount of information and an edge represents direct liking between the two concepts. The similarity among these two concepts is computed on the basis of information they have shared in that space. It assumes if the concepts share more common information, the concepts are more similar. Resnik et al [23] introduced the information content theory in year 1995. In this many resent researches are also taking place presented by different researcher group [24][24], [25].These approaches are corpus dependent it accuracy dependent on the size of the corpus and if same Lcs (least common subsumer) not exist they can't compute the similarity among the words.

Feature based methodology described by the Author Tversky [26], similarity among the concepts as a factor of their properties. It rely upon the extent of common and un-common features of compared concepts. Common features increase the similarity and exceptional features tend to weaken it. It is established on the theory that every concept is depicted through a set of keywords representing their properties and features. As represented in the WordNet through definitions and glosses values. Glossary of the word "automobile" describe through the WordNet as "a motor vehicle with four wheels; usually propelled by an internal combustion engine". Two words are more similar if they contain more common characteristics of words and fewer exceptional characteristics. Tversky measure is represented as;

$$Sim_{Tversky}(C_1, C_2) = \frac{|c_1 \cap c_2|}{|c_1 \cap c_2| + k(|c_1 \setminus c_2|) + (k-1)(|c_2 \setminus c_1|)} \quad (1)$$

K is adjustable factor and K can be $\in [0,1]$

Rodriguez and Egenhofer [27] In this computation is done on the basics of the individual sum of likeness between the "synsets", their "features" and "neighborhood concept" of the estimated terms as;

$$Sim_{Res}(C_1,C_2) = w.S_{synsets}(C_1,C_2) + u.S_{feature}(C_1,C_2) + v.\ S_{neighborhood}(C_1,C_2) \quad (2)$$

$S_{synsets}, S_{feature}$ and $S_{neighborhood}$ are the similarity among the "synonym set" their "features" and corresponding "semantic neighborhood" of the computed terms. According to this (w, u, v>=0) value of w, u and v depends on the particular similarity weights of every specific component. The value relies upon the individuality of the taxonomy. Here S symbolizes the coinciding between the dissimilar features, which are computed by the equ. 1. In the latest research [28] used the Wikipedia as an ontology and similarity is computed on the basic of synonyms, glosses, anchors and categories among the concepts. Main shortcoming of this study is that computational space and huge amount of knowledge is required as for each word; have to search many Wikipedia articles.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

In this research paper, we have utilized the WordNet knowledge i.e. already utilized the textual knowledge about the concepts and computation is low in this case.

III. PROPOSED APPROACH

a. Process flow

The motivation behind this paper is to present a new method in the area of features based similarity measures to tackle the limitation of past methods applied. The current paper focuses on the learning of semantic similarity between the concept / words or terms using fuzzy formal concept analysis, fuzzy membership show how close the word to a particular attribute in a context, as in example the words (car, road, ship) comes in travel ontology but the feature road is more associated with concept car as compared to ship. If a word having more associative with an attribute we define high membership to that attribute as specified by fuzzy set theory given by the researcher Lofti Zadeh [29].We apply the Fuzzy set similarity measure to calculate the similarity among the words inside the ontology. Extracted the features from WordNet a Global Ontology, features usually depict the taxonomic and non–taxonomic information present in ontology. Overall process is shown in Fig.2

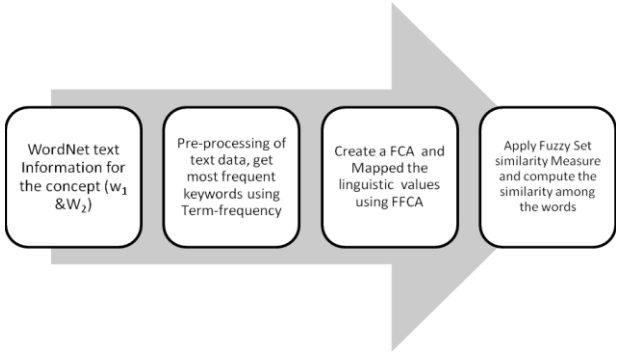


Figure 2. Process flow diagram

B. Preliminaries

Formal concept analysis was introduce by a mathematician Willy et al [30] based on the lattice theory; a group containing same set of objects and attributes. In a defined domain a concept (word) is a set of objects, describes the instances of the concept. FCA is a well established theoretical structure for construction, evaluation and visualization of data for the information used in data mining, knowledge representation and for the development of ontology. FCA based techniques describes formal contexts to symbolize the relationships among the objects and attributes in a given domain. Formal concepts correspond to concept lattice. Concept lattice can be further used for presenting information accurately and effectively[31], [32] .

Fuzzy Formal Concept Analysis (FFCA) proposed by the Author et al Quan in 2004 [33] by introducing the FOGA framework, merging the two theories FCA and fuzzy theory. In this membership value is define among the attributes according to the fuzzy set theory in a particular context. FFCA is a generalization of FCA for defining the degree of fuzzy membership between the object and the attributes. In FCA if a concept is related to some attribute having only two values, if it is related then showing through (☐) otherwise that column is blank, FCA has only two values are (0, 1). FFCA is used for dealing the vagueness and uncertainty present in data. FFCA permits us to quantify the terms how much they are related within a formal context. It shows the closeness between the particular instances with their attributes. "How much" association is illustrated by the membership grade values. For every formal concept a membership value is specified for each attribute values. This information can be used in categorize each formal concept more accurately [34].

Different researchers proposed different similarity measure for fuzzy sets .Fuzzy similarity measures are used to compute the similarity among the fuzzy sets. For this research study, for the similarity computation among the words we have choose a similarity measure introduce by the researcher[35], compute a good accuracy in all fields of application like image processing , fuzzy information retrieval[36] and feature extraction . Fuzzy Similarity between two fuzzy set (A, B) is computed by the given formula.

$$Sim_{(A,B)} = \frac{\sum_{i=1}^{n} min(\mu^i A, \mu^i B)}{\sum_{i=1}^{n} max(\mu^i A, \mu^i B)} \qquad (4)$$

$$\text{If } \mu^i A + \mu^i B = 0 \text{ then } Sim_{(A,B)} = 1$$

IV. PROCESS DEVELOPMENT

Extraction of feature is a difficult task among the ontologies. In this paper, utilized a Global ontology WordNet having gloss, hypernymy / holonymy, antonym and many more relationships of a particular word {w} in WordNet {WN} is present. Using these information, extracted the keyword or the feature /attribute of that particular concept. We construct a multi-feature/attribute row for a single word in the same manner we can construct the multi-features value for another words. Formal definition of a Context; it is comprise of three parameter, one is a concept, other-one are its attributes and last one is the relationship present among the concept and attributes. A formal lattice is drawn using these concepts; one is known as intent and other known as extent for the lattice. It is used for the transforming of data in to lattice. Lattice shows the inheritance relationship "is-a" between the two concepts. Taking an example of travel ontology in WordNet. Let it has only three classes/words {Car, Bus, Ship} which are all automobiles. Random features are represented by {f1, f2 f3….fn} in Formal Concept Analysis. In this, also take an example of another word/class pencil which is not described in above context. Attributes are shown by the features and terms are concepts. Different attributes are depicted in Table 1. for the concepts car, bus, ship and pencil.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Table No-1 Concept/Attribute represented by FCA.

Attribute / Concept	On road	On water	Wheels	Fuel used	price	Fuel consumed	Speed	Produce heat
Car	•		•	•	•	•	•	•
Bus	•		•	•	•	•	•	•
Ship		•	•	•	•	•	•	•
Pencil					•		•	

Drawn a lattice diagram using an open excess tool; Concept Explore[37] for this context. In this, features {speed, price} is common to all concepts, that's why it is most generalized features. In lattice diagram {ship, car, bus} are on same level and pencil which is not part of this context is on another level. Up-to this point we can say class ship, bus, car are more similar to each other as compared to pencil. From the fig. 3, it can be depicted that words bus, car are most similar words but how much similar we can't compute.

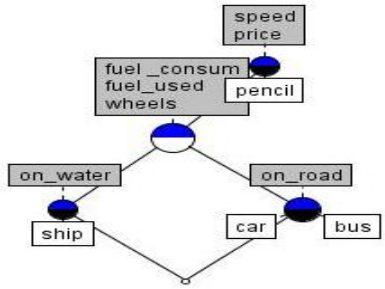


Figure No 3- : Lattice diagram of above example using Concept Explorer Tool

In FFCA fuzzy membership is defined among the objects and attributes. This knowledge can be exploited for measuring the similarity between the formal concepts. In the same example, we can assign the linguistic value to each feature for a particular concept. As we can say that word/concept car is having a high membership towards road as compare to ship having low membership value. We have defined the linguistic value to each attribute/feature according to "how much" it is related to a specific object. Like in word ship has more price as compared to word car and bus, in the same manner linguistic values are defined.

Table No-2: Assigning the Linguistic Values to each feature

Attribute	On -road	On-water	Wheels	Fuel used	Price
Car	High	Very low	Moderate	Moderate	High
Bus	High	Very low	High	Moderate	Very high
Ship	Very low	Very_high	Very high	High	Very very high
Pencil	No	No	No	No	Low

After that we have assign the membership range to each feature. We have used an expert opinion to define the membership values to the features set.

Table No-3: Shows the membership assignment for the different features.

Attribute	On –road	On-water	Wheels	Fuel Used	Price
Car	[0.7-0.9]	[0.01-0.1]	[0.4-0.6]	[0.5-0.6]	[0.5-0.8]
Bus	[0.7-0.8]	[0.01-0.1]	[0.5-0.7]	[0.6-0.7]	[0.6-0.8]
Ship	[0.01-0.1]	[0.8-1]	[0.7-0.8]	[0.7-0.8]	[0.8-0.9]
Pencil	[0]	[0]	[0]	[0]	[0.05-0.1]

Now we can compute the similarity using the fuzzy set similarity measure among any of these two words as (car, bus), (car, ship),(ship, bus) and(car, pencil) by the similarity measure given in the equation(3)

V. EVALUATION AND IMPLEMENTATION

We have taken the information about the concepts (i.e. car, chair) using WordNet Api, where word information, relationship information & gloss information is show in fig. 4. Using R language, pre-processed the textual information. Pre-processing of text-data means removing stop-word for a given language, remove punctuations & stemming process. Using tm, NLP package in R we are able to extract the most frequent word as features of the concepts. In this we select the keywords, whose frequency is greater than 10. Constructed a multi-feature table for the different words using WordNet api.

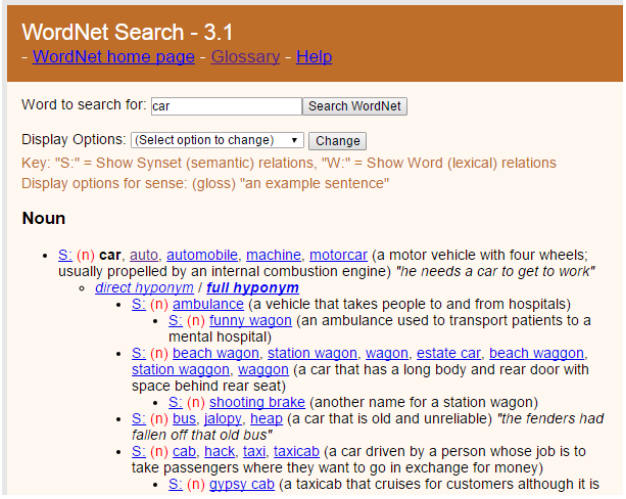


Figure No-4 Word Net Api for the concept "car"

For assigning the membership expert opinion is used. Now we have fuzzy feature vector table (FFVT) for all the concepts, after applying the fuzzy similarity measures we get the final semantic similarity score between the words.

Taking the same example extracted features for word "car". Features are (automobile, travel, wheels, fuel, battery, patrol car, fuel car , speed, machine…….)

Features for the word "bus" (vehicle, public transport, passengers, wheels, ……..).

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Table No-6: FFVT for "Car" and "Bus"

Feature	Road	Water	wheels	Fuel	cost	Speed	travel	Heat_ production
Car	0.96	0.023	0.755	0.623	0.53	0.711	0.93	0.81
Bus	0.84	0.011	0.833	0.73	0.72	0.867	0.754	0.76
Pencil	0	0	0	0	0.01	0.012	0	0

Semantic similarity is computed by equ. 3 as **" *Car* and "*Bus*" is 0.69** and between "***car and pencil***" =**0.012**

$$Sim_{(Car,Bus)} = \frac{0.84+0.011+0.755+0.623+0.53+.711+0.754+0.76+0.31}{0.96+0.023+0.833+0.73+0.72+0.867+0.754+0.81+0.82} =$$

$$\frac{4.534}{6.517} = 0.699$$

Sim(car, pencil)=0.012

VI. RESULT

We have applied this similarity measure on various benchmark dataset like M&C [5], R&G[38] and word-similarity TC-353[39] .Pearson correlation method is used to compute correlation between human judgment and our method.

$$\rho(X,Y) = cov(x,y)/ \sigma X.\sigma Y \qquad (4)$$

Table No-7 Computed Similarity for different words from TC-353 dataset.

Word 1	Word 2	Human (mean)	(Proposed approach)
Love	sex	6.77	0.8
Tiger	cat	7.35	0.9
Tiger	tiger	10	1
Book	paper	7.46	0.8
Computer	keyboard	7.62	0.85
Computer	internet	7.58	0.92
Plane	car	5.77	0.12
Train	car	6.31	0.78
Telephone	communicatio n	7.5	0.87
Television	radio	6.77	0.89
Media	radio	7.42	0.87
Drug	abuse	6.85	0.34
Bread	butter	6.19	0.51
Cucumber	potato	5.92	0.5
Bank	money	8.12	0.89
Wood	forest	7.73	0.85
Money	cash	9.15	0.97

Proposed method has given higher accuracy on different dataset. This method computes the similarity 0.84, 0.83 and 0.82 for the Miller and Charles, R& G and TC-353 dataset respectively. Also compares the results with the other measures in this area.

Table No-8:Comparison Table with other measures.

Meas ures	Type	M&C	R& G	353 word similarity
Path length[40]	Edge –Method	0.59	N/A	N/A
Wu & Palmer[41]	Edge –Method	0.74	N/A	N/A
Lin[19]	IC(Corpus)	0.7	0.72	N/A
Jiang &Conrath[8]	IC(Corpus)	N/A	0.82	N/A
Tversky[2 6]	Feature	0.71	N/A	N/A
Feature based Approach using Wikipedia [28]	Feature	0.78	0.8	0.82
Proposed approach	Feature	0.84	0.83	0.84

VII. CONCLUSION

Semantic similarity judgment is a key issue for the development of semantic web. Semantic web is the future of the current World Wide Web. Now a day's web is used for connecting people and their information's. Semantic web is the extension of Web used to share knowledge among the machines. Ontologies are essential part of Semantic web. To deal with the vagueness and uncertainty among the ontologies, requires the fuzzy ontology. FCA techniques applied in various dimensions as in ontology merging, knowledge discovery and in the field of data mining, without much dealing in the field of semantic similarity. In this paper we have used FCA to compute the similarity among the different words. If two classes/words (that are part of that particular ontology) having the similar features are more related to each other. For similarity computation fuzzy membership is used to describe the proximity towards that attribute. For any two classes (words) having the same attributes and membership towards that attribute are nearly equal or computed more similar word. Through fuzzy formal concept analysis and fuzzy set similarity measure it computes the exact similarity between the concepts/words. Utilizing the word information, we are able to get all the relationship present in Global ontology Word Net. Our experimental result shows that in most cases computed value gives us a better result on various datasets. As this measure is based on Fuzzy formal concept and fuzzy

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

similarity measure this can be further applied for the cross ontologies and in fuzzy ontologies.

This approach can be further modified using neural network for defining the fuzzy membership among the concepts and its feature set.

REFERENCES

1. S. Farrar and T. Langendoen, "A Linguistic Ontology for the Semantic Web," pp. 1–7, 2001.
2. A. Hliaoutakis, G. Varelas, E. Voutsakis, E. G. M. Petrakis, and E. Milios, "Information Retrieval by Semantic Similarity," Int. J. Semant. Web Inf. Syst., vol. 2, no. 3, pp. 55–73, 2006.
3. T. R. Gruber, "A translation approach to portable ontology specifications," Knowl. Acquis., vol. 5, no. 2, pp. 199–220, 1993.
4. B. Dutta, U. Chatterjee, and D. P. Madalli, "YAMO : Yet Another Methodology for large-scale faceted Ontology construction," 2015.
5. G. A. Miller and W. G. Charles, "Language and Cognitive Processes Contextual correlates of semantic similarity Contextual Correlates of Semantic Similarity," Lang. Cogn. Process., vol. 6, no. 1, pp. 1–28, 1991.
6. J. Martinez-Gil, "An overview of textual semantic similarity measures based on web intelligence," Artif. Intell. Rev., vol. 42, no. 4, pp. 935–943, 2012.
7. R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural Language Processing (Almost) from Scratch," J. Mach. Learn. Res., vol. 12, pp. 2493–2537, 2011.
8. J. J. Jiang and D. W. Conrath, "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy," Proc. Int. Conf. Res. Comput. Linguist., 1997.
9. G. a. Miller, "WordNet: a lexical database for English," Commun. ACM, vol. 38, no. 11, pp. 39–41, 1995.
10. A. Budanitsky and G. Hirst, "Evaluating WordNet-based Measures of Lexical Semantic Relatedness," Comput. Linguist., vol. 32, no. 1, pp. 13–47, 2006.
11. C. Naik and P. A. Kharwar, "Knowledge Discovery of Weighted RFM Sequential Patterns With Multi Time Interval From Customer Sequence Database."
12. J. R. Curran, "From Distributional to Semantic Similarity Doctor of Philosophy," p. 177, 2003.
13. R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity," Proc. 21st Natl. Conf. Artif. Intell., vol. 1, pp. 775–780, 2006.
14. A. Aamodt et al., "Requirements Engineering," Requir. Eng., vol. 19, no. 1, pp. 35–46, 2013.
15. B. Xue, C. Fu, and Z. Shaobin, "A study on sentiment computing and classification of sina weibo with Word2vec," Proc. - 2014 IEEE Int. Congr. Big Data, BigData Congr. 2014, no. 2013, pp. 358–363, 2014.
16. A. Hliaoutakis, "Semantic Similarity Measures in MeSH Ontology and their application to Information Retrieval on Medline," Interface, pp. 1–79, 2005.
17. L. Wang and X. Liu, "Knowledge-Based Systems A new model of evaluating concept similarity q," Knowledge-Based Syst., vol. 21, no. 8, pp. 842–846, 2008.
18. K. Saruladha, G. Aghila, and S. Raj, "A Survey of Semantic Similarity Methods for Ontology Based Information Retrieval," 2010 Second Int. Conf. Mach. Learn. Comput., pp. 297–301, 2010.
19. C. Lin and E. Hovy, "From Single to Multi-document Summarization: A Prototype System and its Evaluation," Comput. Linguist., no. July, pp. 457–464, 2002.
20. J. B. Gao, B. W. Zhang, and X. H. Chen, "A WordNet-based semantic similarity measurement combining edge-counting and information content theory," Eng. Appl. Artif. Intell., vol. 39, pp. 80–88, 2015.
21. J. Gao, B. Zhang, and X. Chen, "Engineering Applications of Arti fi cial Intelligence A WordNet-based semantic similarity measurement combining edge-counting and information content theory," Eng. Appl. Artif. Intell., vol. 39, pp. 80–88, 2015.
22. C. Pesquita, D. Faria, A. O. Falcão, P. Lord, and F. M. Couto, "Semantic similarity in biomedical ontologies," PLoS Comput. Biol., vol. 5, no. 7, 2009.
23. P. Resnik, "Using Information Content to Evaluate Semantic Similarity in a Taxonomy," vol. 1, 1995.
24. A. Formica, "Concept similarity in Formal Concept Analysis : An information content approach," vol. 21, pp. 80–87, 2008.
25. A. Banu, S. S. Fatima, K. Ur, and R. Khan, "Information Content Based Semantic Similarity Measure for Concepts Subsumed By Multiple Concepts," vol. 7, no. 3, pp. 85–94.
26. A. Tversky, "Features of similarity. - 1977 - Tversky.pdf."
27. M. J. Egenhofer, "Determining Semantic Similarity Among Entity Classes from Different Ontologies M. Andrea Rodríguez and Max J. Egenhofer IEEE Transactions on Knowledge and Data Engineering," no. 1.
28. Y. Jiang, X. Zhang, Y. Tang, and R. Nie, "Feature-based approaches to semantic similarity assessment of concepts using Wikipedia," Inf. Process. Manag., vol. 51, no. 3, pp. 215–234, 2015.
29. L. A. Zadeh, "Fuzzy Sets," Inf. Control, no. 8, pp. 338–353, 1965.
30. U. Priss, "Linguistic Applications of Formal Concept Analysis," no. 1996, pp. 149–160, 2005.
31. C. Hernández, F. Prieto, M. Laguna, and Y. Crespo, "Formal concept analysis support for conceptual abstraction in database reengineering," Database Manag. Reengineering, no. 34, 2002.
32. U. Priss and L. J. Old, "Modelling lexical databases with formal concept analysis.," J. Univers. Comput. Sci., vol. 10, no. 8, pp. 1–19, 2004.
33. T. T. Quan, S. C. Hui, and T. H. Cao, "FOGA : A Fuzzy Ontology Generation Framework for Scholarly Semantic Web," Knowl. Discov. Ontol., 2004.
34. R. C. Chen, C. T. Bau, and C. J. Yeh, "Merging domain ontologies based on the WordNet system and Fuzzy Formal Concept Analysis techniques," Appl. Soft Comput. J., vol. 11, no. 2, pp. 1908–1923, 2011.
35. V. Cross, "Fuzzy semantic distance measures between ontological concepts," IEEE Annu. Meet. Fuzzy Information, 2004. Process. NAFIPS '04., vol. 2, pp. 1–6, 2004.
36. K. Joshi, A. Verma, A. Kandpal, S. Garg, R. Chauhan, and R. H. Goudar, "Ontology based fuzzy classification of web documents for semantic information retrieval," 2013 Sixth Int. Conf. Contemp. Comput., pp. 1–5, 2013.
37. K. Tunde, J. Rancz, V. Varga, and J. Puskas, "A software tool for data analysis based on formal concept analysis," vol. LIII, no. 2, pp. 67–78, 2008.
38. H. Rubenstein and J. B. Goodenough, "Contextual correlates of synonymy," Commun. ACM, vol. 8, no. 10, pp. 627–633, 1965.
39. A. Budanitsky and G. Hirst, "Semantic distance in WordNet : An experimental , application-oriented evaluation of five measures," Work. WordNet Other Lex. Resour. Second Meet. North Am. Chapter Assoc. Comput. Linguist., vol. 2, no. 12, pp. 29–34, 2001.
40. R. Rada, H. Mili, E. Bicknell, and M. Blettner, "Development and Application of a Metric on Semantic Nets," IEEE Trans. Syst. Man Cybern., vol. 19, no. 1, pp. 17–30, 1989.
41. Z. Wu and M. Palmer, "Verb Semantics and Lexical Selection," 1994.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/335902647

A Review of The Combined Wind and Wave Energy Technologies

Article · May 2019

DOI: 10.14445/22315381/IJETT-V67I5P221

CITATIONS
0

READS
17

4 authors, including:



Prachi Garg
Delhi Technological University
1 PUBLICATION **0** CITATIONS

SEE PROFILE

All content following this page was uploaded by Prachi Garg on 14 December 2019.

The user has requested enhancement of the downloaded file.

A Review of The Combined Wind and Wave Energy Technologies

Dr. J.P. Kesari**, Ashi Gupta*, Kartik Shukla*, Prachi Garg*
** Associate Professor, Mechanical Engineering Department
* B. Tech students, Department of Computer Science and Egg.
Delhi Technological University
Bawana Road, Delhi-110042

ABSTRACT

The efficient development of wind and wave energy systems is a major concern; optimizing resources and developing mechanisms that minimize costs is a pressing challenge. It is with reference to this that we study a combination of wind and wave energy systems that. We will see that these systems offer some exciting avenues in reducing system-wide costs. In the ensuing passages we look at different options of combining existing wind and wave energy systems.

Keywords: *Wind energy, wave energy, floating islands, hybrid systems*

I. INTRODUCTION

A. *Background*

The two forms of energy that dominate the present day ocean energy market are offshore wind and wave [1, 2]. We have been able to develop an understanding of the kind of systems the two incorporate, as well as the potential and the various drawbacks they have to offer. There have been several examples in the recent past where these systems have been effectively employed to harness ocean energy.

Notwithstanding the several advantages promised by these two systems, we still need efficient mechanisms to ensure that they do not overburden the natural resources. It is clear that the development of off-shore wind and wave energy should be done in a manner that does not disturb the ecological balance [3]. For this, one of the primary concerns is related to the efficient utilization of natural resources. Another concern is to bring down the costs involved to make the projects practically feasible. It is in this context that the need for integration arises.

Combining wave and offshore wind energy is a relatively recent approach to tackle the problems highlighted above. The unique synergy effects offer a potential to reduce the cost of electrical energy from offshore units while increasing the quality of the delivered power to the grid. The combination also offers an opportunity to share infrastructures between the systems as well.

Presently, there is a restricted number of articles in the literature that can elaborate on the potential of combined wave and offshore technologies in a sufficient manner. While some of them look to analyze the resource utilization aspects of the hybrid approach [9–13], others look towards integrating the technology to grid systems, generating direct electricity [14, 15].

B. *Objectives*

In the following passages we look at the systems that are necessary to harness the combined potential of the two technologies, the principal advantages and challenges they have to offer, along with a case study on a recent approach developed by Peiffer and Roddier [6].

More precisely, we attempt to analyze the different alternatives for combining wave and offshore wind energy, to give a complete view of the possibilities and current limitations of these systems.

Our document is structured as follows: in section 2 we review the previous work done in this field. In section 3 we discuss the major technologies powering hybrid systems. In section 4 we look at a recent case study, and finally in section 5 we offer conclusions.

II. LITERATURE REVIEW

Some of the earliest work in the field of hybrid technologies was done by Francesco Fusco, Gary Nolan and John V.Ringwood [13]. They presented a methodology to assess the possible benefits of the combination of wind energy with the still unexploited, but quite notable, wave energy in Ireland. Their analysis of the raw wind and wave resources at certain locations around the coasts of Ireland shows how they are very less correlated on the South and West Coast, where the waves are dominated by the presence of high energy swells generated by remote westerly wind systems. They concluded that by integrating wind and waves in combined farms at these locations, it is possible to achieve a more reliable, less variable and more predictable production of electrical power. The resulting benefits are clear especially in the case of a relatively small and isolated electrical system such as the Irish one. In fact, high levels of wind penetration this case strongly increase the requirement of surplus capacity and cause a much lower efficiency for conventional thermal plants.

A seminal work covering different types of avant-garde hybrid technologies was concluded by Pérez-Collazo, C. et al [3]. It presented an ubiquitous analysis of the most relevant aspects related to combined wave and offshore wind energy systems. It developed the different areas related to these systems, aiming to establish a reference frame to be used in future research and developments. The different combined wave and offshore wind systems were reviewed that resulted in a novel classification based on the extent to which the wave and offshore wind technologies are linked. It distinguished three categories, co-located, hybrid, and island systems. Based on this classification, an extensive assessment of the different concepts was presented.

The presence of a combined resource together with the strong synergies existing between both technologies makes a compelling argument for combining wave and offshore wind energy to achieve a sustainable and rational exploitation of the offshore energy resources. Fundamental research was identified as crucial to test the validity, sustainability and integration of these combined systems; to also determine the extent to which these new or adapted WECs are suitable to be combined with current offshore wind farms.

Another work that takes a comprehensive approach is that of [12]. It considers the different aspects concerning the co-location of wave and offshore wind farms and presents a preliminary case study of a hybrid array. As a first step, it presented the issues in synergies and technology development of wave and wind energy systems. In the second phase, different alternatives to combine wave and offshore wind energy systems were presented. Finally, a case study analyzing three possible co-located farms was done. A number of development issues were highlighted as the main challenges for WECs to be integrated with offshore wind. These challenges represent some key research lines and cues which need to be addressed in recent years to realize co-located farms become a reality.

Another noteworthy paper that carried out an extensive analysis of the energy production of a hybrid wave-wind farm was that of [10]. The main objective of this work was to analyze the energy that could be obtained from a hybrid wave-wind farm located in an optimum area for the exploitation of both resources. The authors concluded that such a farm, consisting of offshore wind turbines and co-located Wave Energy Converters (WECs), would be instrumental in realizing the synergies between wave and offshore wind energy. They used data from a hind-cast database spanning 44 years, and investigated the seasonal variability of the hybrid farm's output. They discovered that the hybrid farm indeed constitutes an excellent approach to satisfying the energy requirements of Tenerife and most of the wave and offshore wind energy will be produced in summer mainly because of the high occurrence of winds and waves in the ranges for which offshore wind turbines and WECs are very efficient.

Stoutenburg et. al investigated the power variations of co-located offshore wind turbines and WECs along the California coast [15]. They used Meteorological wind and wave data from the National Buoy Data Center to estimate the hourly power output from offshore wind turbines and WECs at the sites of the buoys. The data set from 12 buoys consisted of over 1,000,000 hours of simultaneous hourly mean wind and wave measurements. It was found that at the buoys, offshore wind farms would have capacity factors ranging from 30% to 50%, and wave farms would have capacity factors ranging from 22% to 29%. Their analysis of the power output indicated that co-located offshore wind and wave energy farms generate less variable power output than a wind or wave farm operating alone. This was due to the low temporal interrelation of the resources and occurs on all time scales. Moreover, their analysis indicated that aggregate power from a co-located wind and wave farm achieves reductions in variability which is equivalent to aggregating power from two offshore wind farms approximately 500 km apart or two wave farms approximately 800 km apart. Combined wind and wave farms in California would have less than 100 hours of no power output per year, compared to over 1000 hours for offshore wind or over 200 hours for wave farms alone. Moreover, ten offshore farms of wind, wave, or both modeled in the California power system would have capacity factors during the summer ranging from 21% (all wave) to 36% (all wind) with combined wind and wave farms between 21% and 36%.

M. Karimirad [16] gauged the feasibility of combining a spar-type offshore wind turbine (inspired by Hywind) and a wave energy converter (inspired by Wavestar) by performing numerical simulations in operational conditions. The dynamic responses and power production of wind turbine and wave energy device are investigated for different power-take-off systems. Coupled/integrated aero-hydro-servo-elastic time domain dynamic simulations considering multi-bodies are applied in this paper. The results show that by choosing proper power-take-off system, it is possible to minimize the effect of WECs on the floating wind turbine and hence maintaining the power performance of the wind turbine while getting more power through wave energy device.

III. TECHNOLOGY

An important factor in combining two different technologies is the resulting synergies they share. This becomes significant in this case, because the combined wind and wave energies offer a great development potential. Concisely, these combined systems can be classified on the following grounds as proposed in [3]:

- the technology used
- required depth of water(shallow, transition, or deep)
- location relative to the shoreline (shoreline, nearshore, offshore)

- degree of connectivity between offshore wind turbines and wave energy converters (co-located, hybrid, and island systems)

Out of the four, we look at the systems defined by the degree of connectivity, primarily because they make up a significant categorization by themselves, as described by Pérez-Collazo, C. et al. [3] in their seminal work.

A. Co-located systems

This is perhaps the simplest, and the most natural option at the current stage of Fundamentally, these systems combine an offshore wind farm with a wave energy converter array. In other words, these systems consist of independent, isolated systems that are brought together. These systems share
several resources like marine area, grid connection, operation and maintenance equipment, etc. Usually,such systems are based on an offshore wind farm, either bottom-fixed or floating. Moreover, these systems do not require any new technology developments.
These system can be further classified into independent arrays and combined arrays.

a). Independent arrays

These systems have separate offshore wind and wave farms, and occupy separate marine areas. But the distance between them is such that they are able to share the same electric grid connection along with other installations. A schematic representation of an independent system is shown as follows.

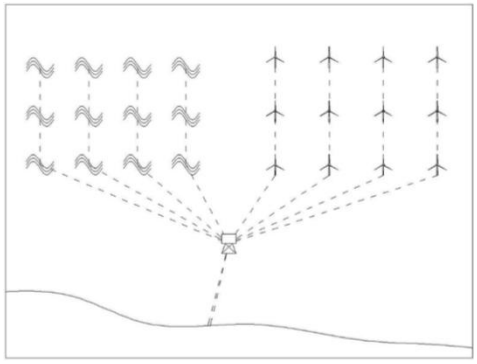


Figure: Schematic of a co-located independent array [3] development of hybrid technologies.

b). Combined arrays

Differing from independent arrays, in this case the offshore wind and wave devices are developed in the same marine area and they share various kinds of infrastructures, in effect, making them behave like single, unified hybrid system. Hence the name single array.

Likewise, combined arrays can be further classified into three categories:

- Peripherally Distributed Array (PDA)
- Uniformly Distributed Array (UDA)
- Non-uniformly Distributed Array (NDA)

The PDA is designed to keep the wave energy converters around the perimeter of the array, in such a way that they follow the direction of the wind, and hence serve as wave shields, protecting the wind devices from the turbulent waves, and consequently, reducing the amount of energy intercepted by the inner section. In a UDA, the wave energy converters are uniformly distributed among the gaps of the wind energy turbines. The NDA does not follow any set patter, and has an uneven distribution of wave energy converters among the wind turbines.

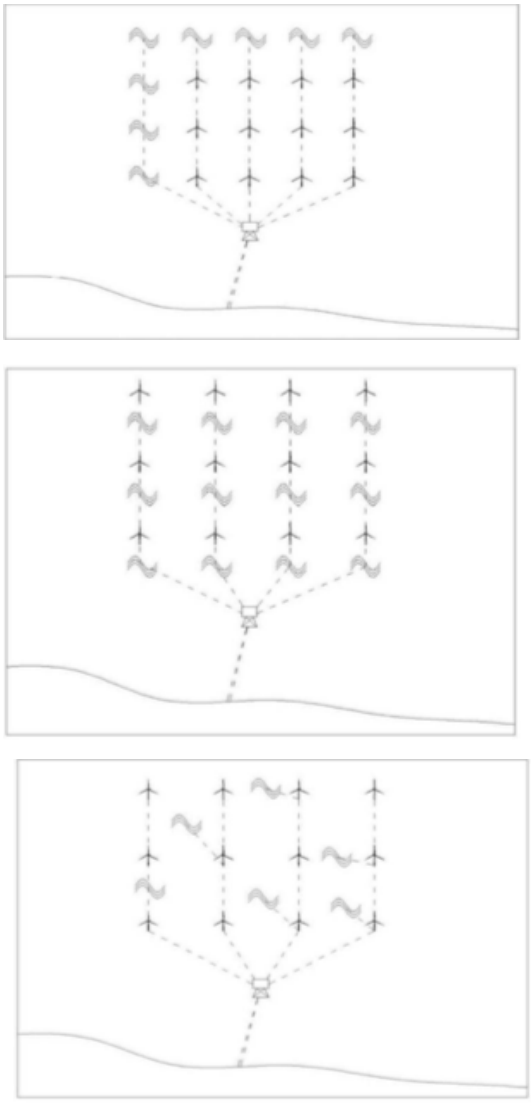


Figure: (From top) Schematic of PDA, UDA, NDA [3]

B. Hybrid systems

A hybrid system combines a wind turbine and a wave energy converter on the same structure.Hybrid systems can be further classified into bottom-fixed, useful for shallow or transitional waters and floating type, appropriate for deep waters.

a). *Bottom-fixed hybrids*

In effect, bottom-fixed hybrids follow naturally from the current systems used by the offshore wind industry. These systems can be modified to accommodate a wave energy converter.



Figure 3: Artist's impression of a bottom-fixed hybrid [3]

Essentially, a bottom-fixed hybrid can be developed by integrating a wave energy converter into an existing offshore wind turbine.

b). Floating hybrids

The floating hybrid is a relatively modern concept that has come into perspective because of the emerging floating offshore wind prototypes, with a considerable number of developers constructing floating substructures that combine offshore wind turbines with wave energy converters.

In the practical world, there have been two EU funded projects, MARINA and TROPOS that dealt precisely with these hybrid systems.

In general, hybrid systems are based on the concept of offshore multipurpose platforms.

C. Island systems

The last category of combined wave and wind energy technologies is island systems. These systems, just like hybrid systems, are based on the ideas of offshore multipurpose platforms. However, unlike hybrid systems, these island systems are usually much larger and integrate the combined exploitation of more than two marine resources on the same platform.



Figure: WindWaveFloat,a floating hybrid wave-wind device (courtesy of Wikimedia [5])

Island systems can be again divided into either artificial or floating islands.

a). Artificial islands

Artificial islands can be developed as platforms for large-scale electricity storage, MRE converters, and other marine activities. A recent and useful concept of this type is that of the Kema Energy Island for Large-scale electricity storage from the Dutch DNV KEMA Consulting.

b). Floating islands

Floating energy islands are large floating multipurpose platforms, and are similar to artificial islands, except on the grounds of capacity and size. These systems are usually of smaller dimensions than artificial islands but larger than most vessels.

IV. CASE STUDY (WIND-WAVE-FLOAT)

A. Introduction

Antoine Peiffer and Dominique Roddier Antoine Peiffer and Dominique Roddier have reported the design of a hybrid device based on combining a wave energy converter on the WindFloat structure. Their seminal paper summarized the numerical modeling and experimental testing that were performed to integrate an Oscillating Wave Surge Converter on the Wind-Float structure [6]. Several further studies have been carried out to investigate the possible solutions by appending different wave energy converters onto the WindFloat device. We look at two such Oscillating water columns and spherical-point absorber devices have been tested.

In general, this hybrid floating wave-wind concept integrating the wave energy converters and the floating WindFloat wind turbine is christened as WindWaveFloat. In design, the WindFloat is a 3-legged floating structure supporting a very large wind turbine, that produces energy greater than 5 MW [6]. The three legs are placed in such a manner that they form an equilateral, and heave plates are fitted at the base of the column, increasing the mass of the total structure. The structure is moored to the seabed, and an electrical cable connects the turbine to the shore. The WindWaveFloat represents an enticing solution for harvesting energy from offshore wind and wave renewable resources.

B. Study considerations

In this section we look at some of the technical details of WindWaveFloat.

In the present stage of floating wind structures, there have been several public demonstrations in the recent past, with different devices being at different stages of development. The HyWind project of Statoil, that features a Siemens 2.3 MW offshore wind turbine has been in use off the coast of Norway since 2009[8]. Principle Power deployed the WindFloat off the coast of Portugal in 2011; featuring a Vestas 2 MW offshore wind turbine.

The WindFloat essentially allows integration with any commercially available, horizontal axis offshore wind turbine. However, other floating support structures may rely on custom turbine designs.

We borrow from the design suggested by [6]. In its basic form, the WindWaveFloat design is based on the power-take-off (PTO) solutions that are directly mounted onto the WindFloat. Using this structure as the basis for an integrated wind-wave energy conversion device offers several opportunities. However, the use of the floating wind support structure as the reference frame for wave PTOs may result in larger motion due to the forces resulting from energy extraction. Therefore, a proper control procedure is needed. One must also optimize the final design to enable an efficient and reliable power production.

Technically, the WindWaveFloat consists of 4 mooring lines, two on column 1, which carries the turbine, and one on each other column. Each line is made of 3 sections: a 3-inch chain at the fair lead, a 5-inch polyester, and another 3-inch chain to the anchor at the bottom. A weight is placed between the upper chain section and the polyester rope to control the tension. The pretension on the mooring lines is 535kN. The displacement of the support structure is 4832 metric tons.

a). Oscillating Water Column

In an oscillating water column (OWC) water enters into a chamber filled with air. As waves pass, the amount of water in the chamber rises and falls like in a piston, thereby sequentially compressing and decompressing the air inside. These positive and negatives changes in pressure result in the ability to direct the air through a bi- directional turbine coupled with an electric generator. A Wells turbine is a common choice for this type of application.



Figure: WindWaveFloat featuring Oscillating Water Column Systems [6]

In this case, the priority is preserving the stability performance of the original WindFloat design, so the chambers are built around the columns that are not responsible for supporting the wind tower and turbine, and for design simplicity's sake extend only 240 degrees around the column, avoiding truss-connecting points on the column.

The effect of the OWC system on support structure motions needs to be understood for both how it might affect performance of wind turbine and how it might affect the performance of the wave energy device.

b). Spherical Point Absorber

The SWEDE (Spherical Wave Energy DEvice) is a spherical point absorber situated in the centre of the structure and attached to the WindFloat columns by three springs and dampening elements. A point absorber is a floating system that absorbs energy in all directions through its movements on the waters surface. A point absorber is usually deigned to resonate so that its harnessed power is maximized.



Figure: WindWaveFloat featuring SWEDE [6]

A spherical floater responds well to heaving with very little pitch motion. The single device SWEDE is a spherical floater installed at the centre of the WindFloat support structure. The floater is attached to the column of the WindFloat by using three lines representing the power take-off system.

C. Economic considerations

By adding a WEC to the structure, the overall economics of the project can be improved. The cost of the WEC alone needs to be less than the energy it produces. However, by sharing both mooring and power infrastructure, the WEC economics become highly advantageous over a farm of WEC alone. WindWaveFloat, along with other innovations in the field of multi-national interlinked offshore-grids, osmotic power projects and offshore energy storage [7] are needed to make a strong impact in the energy space. Ocean energy offers the benefits of significant resources without competing for land use, but blatantly suffers from projected high operating, installation and maintenance costs. The WindWaveFloat concept offers the potential ability to increase the energy production from each floating support structure by approximately

30%. For example, wave energy converters installed in a WindFloat, a floating support structure for large offshore wind turbines, could potentially increase each unit's nameplate capacity from 5 MW to 7.5 MW without a need for additional cabling, space or permits. The ultimate result could lead to the reduction of levelized energy costs by approximately 25%, increase in the overall capacity factor by approximately 10% as compared to the wind only generation, and enhanced stability of the electrical power delivered to the grid.

D. Indian context

The potential of wave energy along the 6000 Km of coast is about 40,000 MW. Primary estimates indicate that the annual wave energy potential along the Indian coast is between 5 MW to 15 MW per meter, thus a theoretical potential for a coast line of nearly 6000 KW works out to 40000-60000 MW approximately. Similarly, India has an estimated 127 GW of offshore wind power potential, mostly off the coasts of Tamil Nadu, Gujarat, and Maharashtra.

Clearly, the two resources abound here, so in terms of brute numbers, combined wave-wind systems have a lot of scope in India. Despite these hopeful signs, we are still far from developing a hybrid offshore system. At the moment, India lacks even the basic infrastructure for a proper offshore wind farm. We do not have a single offshore wind facility that can be combined with existing wave energy conversion systems. Therefore, we have a long way to go before being able to actually implement combined wind-wave systems.

V. CONCLUSIONS

In this paper we have attempted to present an overarching analysis of hybrid wind and wave energy technologies. The main point in combining a wind turbine concept with a WEC concept is to decrease the finished cost of electricity in a long-term perspective. No matter from which side we start, from wave energy or from wind power. The goal is clear which is maturing the offshore renewable energy. The idea of adding a wave-energy device to a floating support structure of a wind turbine is increasing the stability, damping the structure motion by extracting the wave energy and, hence, increasing the produced wind power while taking the wave power, simultaneously. However, as we have seen, the present status of the technology is mostly restricted to prototypical designs. We are still far from pushing this technology into the commercial arena. But the very fact that this sector is less developed than other renewable energy industries presents both opportunities and challenges. The lack of an established industry structure can make entry into the market uncertain for newcomers. However, this lack of structure also means that companies are potentially more able to create and take opportunities than is possible in other parts of the energy industry that are developed and more mature

VI. REFERENCES

[1] Bahaj AS. Generating electricity from the oceans. Renewable Sustainable Energy
[2] Rev 2011.
[3] Iglesias G, Carballo R. Wave energy potential along the Death Coast (Spain). Energy 2009
[4] C.Pérez-Collazo, D. Greaves, G.Iglesias. A review of combined wave and offshore wind energy. Renewable and Sustainable Energy Reviews. 2015
[5] Wavestar. [Online]. Available: http://wavestarenergy.com
[6] Wikimedia. [Online]. Available: wikimedia.org
[7] S. Banister, Alla Weinstein, Dominique Roddier. WindWaveFloat: Combining offshore wind turbines and WEC in a single structure.
[8] The Poseiden Energy website. [Online] Available: www.poseidonenergy.com
[9] The Statoil website. [Online]. Available:
[10] http://www.statoil.com/en/TechnologyInnovation/NewEnergy/Renewab
[11] Veigas M, Iglesias G. Wave and offshore wind potential for the island of Tenerife. Energy Convers Manage.
[12] Veigas M, Iglesias G. A hybrid wave—wind offshore farm for an island. Int J Green Energy 2014.
[13] Veigas M, Ramos V, Iglesias G. A wave farm for an island: detailed effects on the nearshore wave climate. Energy 2014.
[14] Veigas M, Carballo R, Iglesias G. Wave and offshore wind energy on an island. Energy Sustainable Dev 2014.
[15] Fusco F, Nolan G, Ringwood JV. Variability reduction through optimal combina- tion of wind/wave resources—an Irish case study. Energy 2010.
[16] Lund H. Large-scale integration of optimal combinations of PV, wind and wave power into the electricity supply. Renewable Energy 2006.
[17] Stoutenburg ED, Jenkins N, Jacobson MZ. Power output variations of co- located offshore wind turbines and wave energy converters in California. Renewable Energy 2010;
[18] Karimirad M., Koushan Kourosh WindWEC: Combining Wind and Wave Energy Inspired by Hywind and Wavesta

View publication stats

IOP Conference Series: Materials Science and Engineering

PAPER • **OPEN ACCESS**

A Review on Performance and Emission of CI Engine using Exhaust gas recirculation (EGR)

To cite this article: Aseem Dubey *et al* 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **691** 012028

View the article online for updates and enhancements.

This content was downloaded from IP address 14.139.251.100 on 19/12/2019 at 04:47

A Review on Performance and Emission of CI Engine using Exhaust gas recirculation (EGR)

Aseem Dubey[1*], Naushad Ahmad Ansari[2], Girish Kumar[2],Akhilesh Arora[2]andAbhishek Sharma[3]

[1] PG Student, Department of Mechanical Engineering, Delhi Technological University, Delhi, India
[2] Department of Mechanical Engineering, Delhi Technological University, Delhi, India
[3]Department of Mechanical Engineering, GL Bajaj Institute of Technology and Management, Greater Noida, Uttar Pradesh, India

*Corresponding author e-mail:aseemdubey009@gmail.com

Abstract. The diesel engine has the advantage over gasoline engine due to its low fuel consumption, higher combustion efficiency, durability and low hydrocarbons (HC), particulate, carbon dioxide (CO_2), carbon monoxide (CO) and NO_Xemission as a major source. Due to abovesaid problems, diesel is now blended with Biofuels and play a significant role to meet its energy and environmental challenges. Exhaust gas recirculation (EGR) is an eco-friendly alternativeto lower NO_X and particulates emission. The use of EGR in compression ignition engines, lowers the concentration of oxygen mixture and flame temperature. However, it has been observed that advantages of EGR is fuel specific. In present paper a performance review of compression ignitionengine using EGR with blended fuelshas been carried out. The effect on break power, break specific fuel consumption, thermal efficiency has been reported. Further aspects of response surface methodology on optimization of NOx emission, smoke opacity and BSFChasbeen carriedout using blended fuel, EGR rate, injection timing etc.

1. Introduction

Due to depleting problems of fossil fuel resources and harmful exhaust emission of greenhouse gases and NO_X from CI and SI engine, there is need of alternate sources of fuel and new technologies to meet the requirement. In order to growing environmental concern, diesel is now blended with other Biofuels (Dual fuelled s). The use of dual fuel has advantages over CI enginesoperating on diesel fuel alone. Engine operating on dual-fuel, can use various fuels to meet current and future problems. The operation of such engine is based on the interaction of two fuels containing different reactivity level. Biodiesel is produced from both edible and nonedible substances and has advantages in view of current environmental and energy concerns having psychochemical properties and combustion characteristics similar to conventional fuel. It has been reported that use of vegetable oils (i.e. corn, canola,soybean, peanut, rapeseed,jatropha,olive oils, palm oil, coconut and rubber seed) have advantages like low level of sulphur, higher oxygenated nature, higher cetane number and produce less toxic emission[1]. Knothe et al. [2] have reported that vegetable oils have a better lubrication property and a higher ignition temperature than the Diesel fuel and their application are based on

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

geographicaland climate conditions. Song et al. [3] have observed the higher percentage in increase of NO_Xin exhaust gases with increase of oxygen during combustion rather than the use of fuels enriched with oxygen. Lapuerta et al. [4] have found that because of stoichiometric combustion, there is insignificant increase in NO_X emission in biodiesel. In typical biodiesel,stoichiometric combustion found to be occurred at the oxygen-fuel ratio of 2.81,whereas for a standard diesel this ratio is about 3.58.

Various researchers have reported the performance and emission characteristics of engine using biodiesel. It has been observed that reduction in NO_X deteriorates the fuel economy and performance of the engine. It has also been found that comparative thermal efficiency of diesel engine using diesel is higher than the other techniques, but high-level emission of NO_X, smoke and soots are key problems. Some authors [5-7] have reported that with use of biodiesel, the performance of the engine power decreases because of higher viscosity and low heating value. Jinlin *X.* et al. [6] have reported more formation of NO_X in bio diesels as they contain more oxygen content and having higher cetane number. At higher load, engine operates on duel fuel mode show better results in respect ofspecific fuel consumptionand efficiencythan the engine operates on diesel fuel alone, but with increase NO_X exhaust emission. However, at low and medium loads, high specific fuel consumption and lower energy efficiency has been reported. To improve the thermal performance and lowering of NO_X of the dual fuel engine, particularly at low and medium loads, EGR system has been operational. Researchers have found that EGR techniques reducesemission of NOx, as it leads to reduction of oxygen content of the mixture and reduced the combustion temperature below 1371^0C. However, operating points need to be decided by three parameters i.e. % engine load, diesel substitution rate and EGR rate.

As the combustion of mixture with the use EGR is not happened properly, causes formation of excess soot. The higher soot emission leads to degradation of lubricating oil, more carbon deposits, and enhanced wear on engine parts. It has been reported that due to high specific heat of CO_2in EGR, peak cylinder temperature gets reduced. Due to better combustion, rate of pressure rise at various loads is higher if hot EGR system used, comparedto cooled EGR technique. Divakar et al. [8] have carried out Real-time model computation to analysed the attainment of ultra-low NO_X emissions at various loads and boost set-points at controlled EGR effectiveness. Daisho Y et al. [9] have reported that with the use of cooled EGR, the combustion noise and NOx % may be decreased. Walke et al. [10] have found with higher EGR, the NO_X level decreases significantly with slight increase in BSFC and smoke opacity.

The operating conditions for particular fuel with optimized EGR rates may not fit well with the other fuel because of change in chemical properties of each fuel undersimilar working conditions.Kumar et al. [11] reported the effect of increase in compression ratio (CR) and found inincreasebrake thermal efficiency and decrease in specific fuel consumption. As the EGRpercentileincreases, at different compression ratio, there is decrease in pressure rise rate and thermal efficiency, whereas NO_X emission increases from 11% to 85%. Mohamed [12] has reported that to reduce the combustion noise and emission of NO_X, 5% EGR is favourable. Kumar et al. [13] have reported that fuels like gasoline, di-methyl carbonate (DMC) and alcohols in CI engines produced difficultybecause of low cetane number, low calorific value, high volatility and latent heat of vaporization. However, these fuels can be used suitably when blended with diesel [14]. Saleh [15] has found that with the use of 5% EGR at part load, the performance of LPG-diesel engine (with 70% propane in LPG)improved. Poonia and Mathur [16] have studied the effect of EGR and load on LPG-diesel engine and observed that variation in peak pressure and indicated mean effective pressure in cycle are lower up to 60% of load compared to engine operating without EGR. Poonia and Mathur [17] have reported the insignificant effect of hot EGR on combustion delay whereas cold EGR increases. Rajan et al. [18] have carried out the study using sunflower methyl ester as fuel using EGR technique on a DI engine. They observed that 15% EGR reduced the NO_X emission and smoke by

25%. Wagner et al. [19] have used highly diluted intake mixture and found that at 44% EGR, there is sharp decrees in PM emission with a drop in NO_X, however affect the fuel economy significantly. Vallapudi et al. [20] have carried out combined effect of EGR and injection pressure on NO_X emission using tamarind seed biodiesel blend. They have reported that 10% EGR at 220 barslower the NO_Xemission by 80.5% of standard conditions with insignificant decrease in performance.Kumar and Kumar [21] have reported the effect of EGR using Jatropha oil methyl ester as the pilot fuel for compressed natural gas on dual fuel engine with variable load and reported improve in performance and emissions characteristics, at low and intermediate loads. Raju and Kishore have reported 45.67% reduction of NO_X with tamarind seed methyl ester (TSME 20) at 20% EGR as compared to diesel fuel with TSME 20 (52.69%). However, there is a slight reduction in brake thermal efficiency [22].

Performance of an CI engine with EGR may be enhanced by optimizing the design and operating parameters. In this review, influence of exhaust gas recirculation on combustion characteristics and performance of diesel engine (both single and dual fuel mode) has been carried out. The comparative application of EGR on blends of high n-butanol, n-pentanol with diesel, Palm-Biodiesel, Cottonseed Biodiesel fuel etc.at various conditions have also been reprted.

2. Classification of EGR system

Figure 1 shows the schematic of EGR system. EGR systems have been classified as under;

2.1. *Based on temperature*
(a) Hot EGR (b) Partly cooled EGR and, (c) Fully cooled EGR.

2.2. *Based on configuration*
(a) Long route and, (b) Short route.

2.3. *Based on pressure*
(a) Low pressure system and, (b) High pressure system.

In hot EGR cycle, some exhaust gas recycled back to intake directly. Whereas, in cooled EGR system re cycled exhaust gas is pre-cooled. The engine using hot EGR promote better combustion with increase in thermal efficiency. The volumetric efficiencyincreases with cooled EGR with decrease of NOx, however unburnt hydrocarbons increases in exhaust.

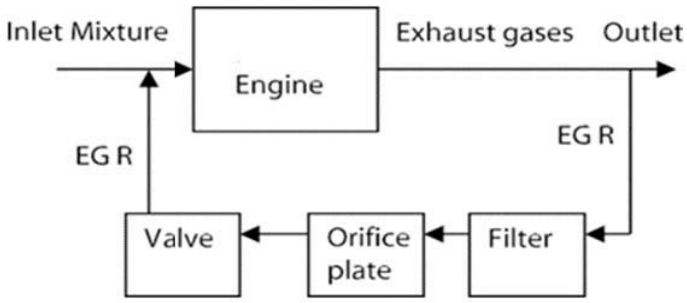


Figure 1. Schematic of EGR Diesel Engine.

The % EGR is calculated using the Eq. (1)

$$EGR = \frac{Volume\ of\ EGR}{\textbf{Total}\ intake\ charge\ into\ cylender} \times 100 \quad (1)$$

3. Performance and emission characteristic of EGR engine

Performance of CI engine using EGR is influenced by speed, percentile of blend, load, fuel quantity, injection timing, fuel compositions, engine speed, SOI advance, EGR % etc. The results presented by some authors as effective measures for furtherimprovement in performance taking into various parameters, are as under;

3.1 Effect of EGR on diesel fuel

Agrawal et al. [23] carried out experimentation on air cooled two-cylinderdiesel engine of rpm 1500using diesel fuel, EGR rates (0-20%) andobserved that:

(i) NO_X emission reduced substantiallyat 15% EGR without deteriorating the BSFC,thermal efficiency,and emissions.
(ii) Increase rate of EGR reduces NO_X. However, performance decreases at higher loads. The effects of EGR change on thermal efficiency and BSFC is insignificant at higher engine load.
(iii) There ismore wear of piston rings due to higher carbon deposits on the engine parts.

Chowdary et al. [24] carried out experimentation on constant speed single cylinder C.I, with different EGR rates (0-10%) and advanced start of ignition (SOI) from 11^o bTDC to 14.5^o bTDC and following have been observed:

(a) Effect of EGR (10%) with SOI 11 degree
NO_X increases significantly with increase in crank angle from 0-32^o and remains nearly constant further till 150^o crank angle. SOI 11^o& EGR 10% has reduced the emission of NO_X by 34.4%. whereas there is decrease in cylinder peak pressure about 0.30 MPaand increase in soot by 37.2%.

(b) Effect of SOI only (11-14.5 Degree with EGR 0%
NO_X increases with 7.4% with increase in SOI advanced by 3.5 degree. Soot increases by 37.2% but decreases at the end of combustion about 52% with advanced SOI. Further there is increase in peak cylinder pressure with advanced engine.

(c) Effect of SOI 14.5^o with EGR 10% compared to SOI 11^o with 0 EGR
NO_X decreases about 1.2% at the end of combustion, Peak pr. increases by 0.8 MPa and decrease of soot at the exhaust by 21.3%.

3.2 Effect of EGR on Neat palm-bio diesel blend

Yasina et al. [25] carried out experimental analysis using neat palm-bio diesel blend on water cooled diesel engine (four stroke),operating atvariable 1000-3000 rpm and observed that;

(i) Brake power increases with RPM up to 2500 rpm. However, the effect with B100 is less with increase EGR as compared to Diesel fuel. Further, torque decreases with increase in RPM and with B100 with increase EGR.
(ii) BSFC increases with rpm and EGR and higher than Diesel. Emission of CO, CO_2 and UHC increases with RPM and higher with B100 than Diesel.
(iii) With increase in engine speed, NO_X emission is less using EGR whenoperating with B100 fuel than the use of diesel only.

3.3. Effect of EGR on Cotton seed Bio diesel blend

Sakharea et al. [26] carried out an experimental work, using cotton seed Bio diesel blend on single cylinder fixed rpm (1500) CI engine with variable EGR (4-6%). It has been observed that;

(i) Biodiesel (B20) an oxygenated fuelhas shorterignition delay as compare to diesel due to higher cetane number. The decreased in ignition delay has been recorded from 8 ºCA to 7 ºCA with 6% of EGRcompared to diesel fuel alone.
(ii) With use of Diesel fuel, BSFC increasesfrom 7.69 % (4 % EGR) to 11.53 % (6 % EGR) compared to use of diesel fuel alone. Whereas there is increase of BSFC from 30.76 % (4 % EGR) to 42.60 % (6 % EGR) using B20 biodiesel in comparison to diesel fuel.
(iii) At the rated load, thermal efficiency decreases from 1.33 % (4 % EGR) to 3.9 % (6 % EGR) in comparison to use of diesel fuel alone. Whereas there is decrease in thermal efficiency from 4.95 % (4% EGR) to 5.1 % (6 % EGR) using B20 biodiesel in comparison to diesel fuel.
(iv) Decrease in emission of COand UHC is higher with increase EGR using B20 biodiesel in comparison to diesel fuel. However, NO_X level in exhaust is higher in comparison to base diesel and decreases with increase in EGR.

3.4. Effect of EGR on diesel-chaulmoogra oil - diethyl ether blend

Krishnamoorthi et al. [27] reported experimental work on variable speed CI compression ignition engine at 1500-2100 rpm, using diesel-chaulmoogra oil diethyl ether blend, EGR varies from 0, 5 & 10% and optimization using ternary blend i.e. for 65% diesel+ 25% straight vegetable oil + 10% diethyl ether blend. It has been observed and concluded that:

(i) Optimum values of the blends determined as 0% EGR (80% load), 5% EGR (77% load) and 10% EGR (73% load). maximum thermal efficiency of 28.85% has been observed with 10% EGR, 20.6 compression ratio and 1500rpm with 100% load. However minimum BSFC (0.52 $kgkW^{-1}$ hr^{-1}) has been found at compression ratio of 19.2and at 1800 rpm,at full loadand 5%EGR.
(ii) At high compression ratio and speed at full load with 10% EGR, engine produced maximum smoke with higher NOX emission at 1500 rpm. Maximum UHC emission of 47.7 ppm has been observed at compression ratio of 17 and 1500 rpm at 25% engine load with 10% EGR.

3.5. Effect of EGR on diesel and CNG

Selim [28] reportedeffect of EGR (5-15%) on single cylinder indirect injection CI engine with variable compression ratio and speed (1200-2400 rpm)as under:

(i) Efficiency increases with increase in the speed from 1200 - 2400 rpm. Increase in efficiency is less significant with increase in EGR ranges up to 5% for low speed (1600 rpm). It has been further observed that with increase in EGR from 10 and 15%, the thermal efficiency tends to decrease irrespective of engine speeds.
(ii) 5% EGR is favourable toincrease the thermal efficiency, reduction of combustion noise and NO_X emission.

4. Optimization of Exhaust Gas Recirculation System

Response Surface Method is a useful technique, for optimizing output responses obtained with the change of fuel, EGR, injection timing etc. The experimental data recorded are further analysed using regression and polynomial models. In order to solve multi response problems and to get optimum combination RSM along with the desirability approach is used effectively with blended fuel with reasonable accuracy. The optimization effects reported by some researchers using EGR for various responses are:

Kumar et al. [29] carried out experimental work using three types of blends namely dimethyl-carbonate (DMC), isobutanol and n-pentanol on direct injection CI engine, varying the EGR as 0, 15, 30%, high-load condition and injection time change 21, 23. 25^o CR bTDC. In order to minimize

$NO_X$and smoke emission and simultaneously BSFC, the EGR rate had been optimized with injection timing and observed that;

(i) The point of high EGR rate and small injection timing is best suitable condition to minimize NO_X emissions and found to be higher with the use of n-pentanol blend with diesel compared to blends of Isobutanol/diesel and DMC/diesel. Further with retard of injection timing from 25^o to 21^o bTDC and increase EGR rate, NO_X emission decreases possible due to lower combustion temperatures irrespective of blends.
(ii) The advanced injection timing and low EGR rate produce best smoke opacity. Smoke opacity decreases to grater extant with blend of Iso butanol and diesel in comparison to use of blends of DMC/diesel and n-pentanol/diesel.
(iii) The BSFC was found lower for all fuel blends under study and observed minimum with the engine operating with blend of isobutanol/diesel
(iv) The optimum combination is obtained with blends of isobutanol/diesel compared to diesel alone, injected at 22^o bTDC under natural condition. The reduction in 12.4% NO_X and 100 % smoke opacity with insignificant increase in BSFC (2.9%) has been found to be optimum.

Praveena et al. [30] experimentally tested diesel engine with Mahua oil and its blends with varying EGR between 10 to 30%, injection pressures (200 & 230 bar), engine load and optimum EGR for 20% Mahua oil blend. It has been observed that:

(i) There is nearly equal brake thermal efficiency using Mahua oil blend of 20% at pressure of 200 bar, with reduction of UHC, CO and carbon particulate in exhaust. However, produces more concentration of nitrous oxide gasesthan the use of diesel as a base fuel alone. The aforesaid condition is observed asgood agreement in view of specific fuel consumption as well as BTH efficiency.
(ii) With increase of blendproportion, emission of nitrous oxide gases increases, however reduced at higher loads.

Bhaskar et al. [31] carried an experimental study on single cylinder Diesel engine using Fish oil methyl esters (FOME) and its blends, EGR 10,20, 30% at 20% blend, constant RPM of 1500 and Diesel, FOME & their 20-80% blend were used. It has been observed that;

(i) the brake thermal efficiency is found to be lower all brake power outputs, compared to diesel fuel and observed to be decreased with increase in blend %.
(ii) At all loads, emission of CO, UHC and soot using FOME/ blends is lower, which decreases further with increase in blend percentage. The optimum value brake thermal efficiency has been obtained with 20% blend in comparison to diesel.
(iii) Percentage increase in NOx emission has been reported with increases of FOME %. NO_X emissions were decreased by employing exhaust gas recirculation and reported optimum with 20% blend at 20% EGR rate considering emission of NOx and soot.

Gaikwad et al. [32] reported an experimental study on water cooled single cylinder diesel engine using blend of diesel/LPG at varying EGR rate (i.e. 5%, 10% and 15%) and load (25 to 100%). It has been observed that at 25% load, change in BTE is insignificant. At 50% to 75% load, a slight increment in the BTE is obtained with EGR and it increased with EGR rate. Slight reduction in BTE is observed with EGR at full load and it decrease with EGR rate but higher than straight diesel operation.

Comparative study has been performed on two duel fuel diesel engines at full load conditions, one operates with LPG (LPG of quantity 60% energy share is fumigated with intake air stream at ambient pressure and the diesel is injected by traditional way with injection) and another with Mahua oil blend of concentration 20% in EGR at injection pressure of 200 bar with view point to compare

NO_Xemission and BTE. The comparative results are plotted in Fig. 2. It has been found that there is significant decreasein the emission of nitrous oxide gasesusing Mahua blend compared than the use of LPG duel fuel engine. However, decrease of NO_X with EGR is not much significant with the use of Mahua blend as compared to LPG fumigated engine. This may be due to lean mixture of blend produces low-combustion temperature. Emission of nitrous oxide gasesNO_Xdecreses significantly using EGR in duel fuel operation than the use of strait diesel fuel alone (970 ppm).

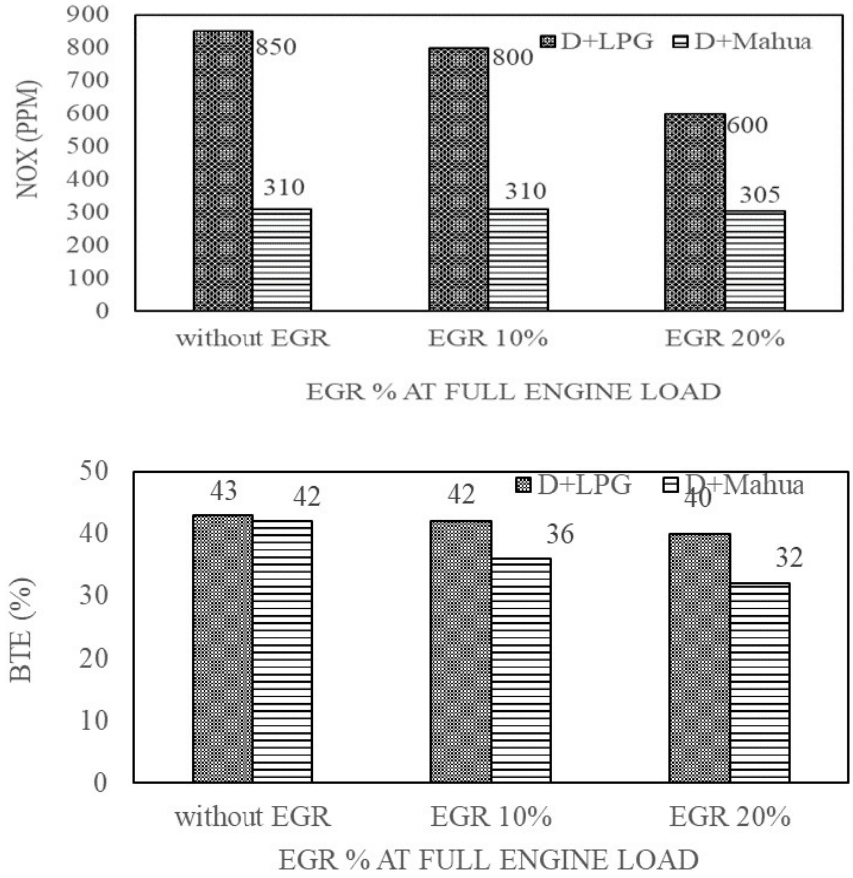


Figure 2. Comparative variation of NO_X emission & BTE using duel fuel with EGR at full Load [27, 29].

Break thermal efficiency with use of Mahua blend has been found to be lower than CNG operated engine and decreases further with increase in EGR rate. Higher efficiency has been foundif engine operates on dual fuel without using exhaust gas recirculationin comparison to use of dieselas a fuel (about 33%). This is because for higher loading condition, pilot fuel quantity increases and it led to better combustion of LPG fuel. It has also observed that BTE with Mahua blend is almost same at 20% EGR, may be due to more fuel supplied in the cylinder at higher load. However, with EGR, combustion temperature is lowered because of reduced oxygen concentration as a part of O_2 get replaced by CO_2 and some H_2O.

5. Conclusions

The following results have been summarised from the review.

1. Using strait diesel as a fuelat lower loads incorporating EGR technology, there is insignificant improvement in break thermal efficiency with decrease BSFC. However, the said efficiency and BSFC are almost remain same at higher loads. Therefore, at lower load higher EGR rate is more advantageous.
2. The NO_X emission increases and soot decrease with advanced SOI. Therefore, combined effect SOI advancement and EGR is useful to reduceemission of nitrous oxide gases and carbon particulates. There is increased work output/IMEP and efficiency with combined effect.
3. Dual fuel mode without EGR gives lower emissions of NO_X than straight diesel engine. When the EGR is used in dual fuel mode NO_X emission decreases significantly and it decreases further with increasing EGR rate. The lower oxygen concentration decreases the peak pressure rise rate of pre-mixed combustion with decrees of combustion temperatures and pressures and, consequently, to lower NO_Xbut increase of shoots.
4. Torque decreases with increase in RPM and EGR simultaneously decrease in Pressure rise rate.
5. BSFC increases with rpm and EGR and higher than Diesel fuel in other blended fuels.
6. Optimum performance with considerable agreement between soot and NO_X emissions is obtained at 20% EGR rate with 20% blend at 100% load. However, variation is fuel specific.
7. NO_X emission decreases with retard of injection timing from 25^o to 21^oCA bTDC and increase EGR proportion. Optimum emission of Nitrous oxide gases occurred with high EGR rate and retarded injection timing. Effect is observed significantly higher using blends of n-pentanol compared to Iso butanol and DMC in base diesel.
8. There is no majormodification of CI engine required once it operates on Mahua oil and its blends. However, with increase percentage of Mahua oil, the brake thermal efficiency decreases correspondingly and found to be optimum at 20% blend with 20% EGR.

6. References

[1] Heng Sit Foon CYM, Yusof Basiron, Yung Chee Liang and Ma Ah Ngan 2005, Palm Biodiesel: Gearing Towards Malaysian Biodiesel Standards, Malaysia Palm Oil Board (MPOB) 28-34.
[2] Knothe G, Gerpen JV and Krahl J.2005, The Biodiesel Handbook: American Oil Chemists' Society Press.
[3] Song JVZ and Boehman AL 2004, Comparison of the impact of intake oxygen enrichment and fuel oxygenation on diesel combustion and emissions, Energy and Fuels 18 1282-90.
[4] Magin Lapuerta OA and Fernandez JR 2007 Effect of biodiesel fuels on diesel engine emissions *Progress in Energy and Combustion Science* **34** 198–223.
[5] Rizwanul IM, Masjuki HH, Liaquat AM, Ramli R, Kalam MA and Riazuddin VN 2013 Impact of various biodiesel fuels obtained from edible and non-edible oils on engine exhaust gas and noise emissions Renewable and Sustainable Energy Reviews 18 552–567.
[6] Jinlin X, Tony E and Hansen A 2011 Effect of biodiesel on engine performances and emissions *Renewable and Sustainable Energy Reviews***15** 1098–1116.
[7] Jaichandar S and Annamalai K 2011 The status of biodiesel as an alternative fuel for diesel engine *Journal of Sustainable Energy & Environment***2** 71-75.
[8] Divekar P, Xiang QT and Zheng CM 2015 Characterization of Exhaust Gas Recirculation for Diesel Low Temperature Combustion IFAC-Papers on Line 48-15 045–051.
[9] Daisho Y, Yaeo T, Koseki T, Saito T, Kihara R and Quiros EN 1995 Combustion and Exhaust Emissions in a Direct-Injection Diesel Engine Dual-Fuelled with Natural Gas SAE Paper No. 950465.
[10] Walke DNVD PV, Bodkhe RG 2008 Impact of Exhaust Gas Recirculation on the Performances of Diesel Engine Proceedings of the World Congress on Engineering (WCE) London U.K.
[11] Kumar NR, Sekhar YMC and Adinarayana S 2013 Effects of compression ratio and EGR on

performance, combustion and emissions of DI injection diesel engine International Journal of Applied Science and Engineering 11 (1) 41-49.

[12] Mohamed YES 2003 Effect of exhaust gas recirculation on some combustion characteristics of dual fuel engine *Energy Conversion and Management***44** 707–21.

[13] Kumar B Rajesh, Saravanan S, Rana D, Anish V and Nagendran A 2016 Effect of a sustainable biofuel – n-octanol – on the combustion, performance and emissions of a DI diesel engine under naturally aspirated and exhaust gas recirculation (EGR) modes *Energy Conversion and Management* 118 (2016) 275–86.

[14] Kumar B Rajesh, Saravanan S, Rana D and Nagendran A 2016 Use of some advanced biofuels for overcoming smoke/NOx trade-off in a light-duty DI diesel engine Renewable Energy 96 687-99.

[15] Saleh HE 2008, Effect of variation in LPG composition on emissions and performance in a dual fuel diesel engine Fuel 87 3031-39.

[16] Poonia MP and Mathur YB 2012 Effect of EGR injection rate and engine load on cyclic variations of an LPG diesel dual fuel engine *Journal of Environmental Research and Development***7** 746-55.

[17] Poonia MP and Mathur YB 2012 Effect of Exhaust Gas Recirculation on the Combustion of an LPG Diesel Dual Fuel *Engine Engineering Science and Technology* 2 2250-98.

[18] Rajan KRS K 2009 Effect of Exhaust Gas Recirculation (EGR) on the Performance and Emission Characteristics of Diesel Engine with Sunflower Oil Methyl Ester Jordan Journal of Mechanical and Industrial Engineering 3 306 -11.

[19] Wagner RM, Green Jr JB, Dam TQ, Edwards KD and Storey JM 2003 Simultaneous low engine-out NOx and particulate matter with highly diluted diesel combustion SAE 2003-01-0262.

[20] Vallapudi DR, Makineni HK, Pisipaty SK and Venu HC 2018 Combined impact of EGR and injection pressure in performance improvement and NOx control of a DI diesel engine powered with tamarind seed biodiesel blend Environmental Science and Pollution Research 25(36) 36381-93.

[21] Kumar P and Kumar N 2016 Effect of EGR on performance and emission characteristics of a dual fuel engine fuelled with CNG and JOME Biofuels 7 743-51.

[22] Raju VD and Kishore PS 2018 Effect of exhaust gas recirculation on performance and emission characteristics of a diesel engine fuelled with tamarind biodiesel International Journal of Ambient Energy 1-10.

[23] Agarwal D, Singh SK and Agarwal AK 2011 Effect of Exhaust Gas Recirculation (EGR) on performance, emissions, deposits and durability of a constant speed compression ignition engine *Applied Energy***88** 2900–07.

[24] Chowdary PK, Ganji PR, Kumar MS, Kumar CR and Rao SS 2016 Numerical analysis of C.I engine to control emissions using exhaust gas recirculation and advanced start of injection Alexandria Engineering Journal 55 1881–1891.

[25] Yasina HM, Mamata R and Yusopa AF, Parukaa P, Yusaf T and Najafic G 2015 Effects of Exhaust Gas Recirculation (EGR) on a Diesel Engine fuelled with Palm-Biodiesel Energy Procedia 75 30 – 36.

[26] Sakharea NM, Shelkea PS and Lahanea S 2016 Experimental Investigation of Effect of Exhaust Gas Recirculation and Cottonseed B20 Biodiesel Fuel on Diesel Engine *Procedia Technology***25** 869 – 76.

[27] Krishnamoorthi M, Malayalamurthi R and Sakthivel R 2018 Optimization of compression ignition engine fuelled with diesel - chaulmoogra oil - diethyl ether blend with engine parameters and exhaust gas recirculation, Renewable Energy, 0.1016/j.renene.2018.11.062.

[28] Selim YE 2003 Effect of exhaust gas recirculation on some combustion characteristics of dual fuel engine Energy Conversion and Management 44 707–721.

[29] Kumar B. R., Saravanan S, Rana D and Nagendran A 2016 Combined effect of injection timing

and exhaust gas recirculation (EGR) on performance and emissions of a DI diesel engine fuelled with next generation advanced biofuel – diesel blends using response surface Methodology *Energy Conversion and Management***123** 470–486.

[30] Praveena V, Shivam R, Ronith 2017 optimization of mahua biofuel blend and EGR (exhaust gas recirculation) for control of Nox emissions in diesel engines *Jr. of Industrial Pollution Control***33(2)** 1670-1675.

[31] Bhaskar K, Nagarajan G and Sampath S 2013 Optimization of FOME (fish oil methyl esters) blend and EGR (exhaust gas recirculation) for simultaneous control of NOx and particulate matter emissions in diesel engines *Energy***62** 224-234.

[32] Gaikwad M, Jadhav K., Kolekar A 2016 Experimental Analysis of Dual Fuel Compression Ignition Engine with Exhaust Gas Recirculation, International Journal of Current Engineering and Technology Special Issue-5 480-486.

arXiv:1912.01667v1 [cs.LG] 3 Dec 2019

A Study of Black Box Adversarial Attacks in Computer Vision

SIDDHANT BHAMBRI* and SUMANYU MUKU*, Delhi Technological University, India
AVINASH TULASI and ARUN BALAJI BUDURU, Indraprastha Institute of Information Technology Delhi, India

Machine learning has seen tremendous advances in the past few years which has lead to deep learning models being deployed in varied applications of day-to-day life. Attacks on such models using perturbations, particularly in real-life scenarios, pose a serious challenge to their applicability, pushing research into the direction which aims to enhance the robustness of these models. After the introduction of these perturbations by Szegedy et al. [1], significant amount of research has focused on the reliability of such models, primarily in two aspects - white-box, where the adversary has access to the targeted model and related parameters; and the black-box, which resembles a real-life scenario with the adversary having almost no knowledge of the model to be attacked. We propose to attract attention on the latter scenario and thus, present a comprehensive comparative study among the different adversarial black-box attack approaches proposed till date. The second half of this literature survey focuses on the defense techniques. This is the first study, to the best of our knowledge, that specifically focuses on the black-box setting to motivate future work on the same.

CCS Concepts: • **Security and privacy → Software and application security**; • **Computing methodologies** → *Computer vision*; *Machine learning*.

Additional Key Words and Phrases: Deep neural networks, Deep Learning, Adversarial examples, Computer vision

ACM Reference Format:
Siddhant Bhambri, Sumanyu Muku, Avinash Tulasi, and Arun Balaji Buduru. 2019. A Study of Black Box Adversarial Attacks in Computer Vision. 1, 1 (December 2019), 32 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Machine Learning (ML) is the most influential field in sciences today. With applications in Computer Science, Biomedical Engineering, Finance, and Law leveraging the edge in Language, Vision, Audio with specialized branches like Deep Learning (DL), Reinforcement Learning (RL) the involvement of ML in day-to-day life is undeniable. RL models are beating world champions at their own game, DL models can classify a multitude of categories with precise identification of species, objects, material which is humanly impossible. The field has grown so much that Deep Language models are writing fiction, news reports, and poetry at the same time. While the developments in the field

*Both authors contributed equally.

Authors' addresses: Siddhant Bhambri, siddhantbhambri_bt2k16@dtu.ac.in; Sumanyu Muku, sumanyumuku_bt2k16@dtu.ac.in, Delhi Technological University, Rohini, India; Avinash Tulasi, avinasht@iiitd.ac.in; Arun Balaji Buduru, arunb@iiitd.ac.in, Indraprastha Institute of Information Technology Delhi, Govindpuri, India.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2019 Association for Computing Machinery.
XXXX-XXXX/2019/12-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

are exciting, the underlying techniques are probabilistic making the performance of an ML model highly dependent on the training data.

Gaps in the training data has lead to biased models, a biased model can have excellent performance on a set of inputs closer to the training data but it fails on corner cases where the input was rarely seen while training. Whatever be the domain, an ML model draws boundaries between classes in an n dimensional space, and these decision boundaries can not be reasoned. Sometimes these boundaries can be subtle, making the model categorize inputs into different classes (with high confidence scores) in the immediate neighborhood of a boundary. Taking advantage of these traits in ML, to fool the system into wrongly interpreting an input, Adversarial attacks are crafted.

A typical Adversarial attack starts with studying the model, inputs, outputs, training data, model parameters etc. according to the threat model. While, earlier threat models were surrounded around the training data and complete access to the model parameters, it is less practical in a sense that the attacker can be an outsider. Black Box adversarial attack on the other hand limits the attacker's capabilities, with access only to the deployed model, sometimes restrictions are also put on the observable parameters of the model. The setting of Black Box attacks and constraints on attacker are closer to the real world scenario.

To understand the Adversarial attacks in the context of Black-box in this section we introduce the Taxonomy.

1.1 Taxonomy

Here we introduce the key concepts in the Adversarial attacks context.

- **Adversary:** An entity trying to attack the Machine Learning model to wrongly classify a perfectly legitimate looking input. Say input $i \in c$, where c is the original class, Adversary's goal is to make the model predict an input $i` \in c`$, where $i`$ is the input adversary carefully crafts and the point to note is $i`$ will still be classified as c for a human annotator.
- **Perturbation:** To achieve the above mentioned goals Adversary crafts an input $i`$ which is perturbed from $i$ by adding an impurity say $\delta\grave{\epsilon}$. Modified input $i` = i + \delta\grave{\epsilon}$ is called the perturbed input. Constraints on the perturbation are a challenge to the Adversary. Any changes made should be under a certain value, to make sure that the input is still classified into its original class c by a human annotator. $L2$ distance is often used in literature to define the acceptable impurity.
- **Adversarial Input:** Input crafted by Adversary after perturbation is called Adversarial Input. Key properties of this input is the impurity and actual input. Addition of this impurity makes the actual input "jump the classification boundary", hence the model mis classifies an Adversarial Input.
- **Targeted Attack:** When the target mis-classified class $c`$ of a perturbed input $i`$ is defined then it is a Targeted Attack. Here, the adversary targets the impurity in such a way that the model classifies given input i to class $c`$.
- **Query:** A query is counted as a single instance of sending input the the model under attack and noting the observations made. Minimizing the number of queries made reduces the time taken to build an adversarial example, hence it is a key aspect of building efficient adversary models.
- **Threat Model:** Threat model defines the rules of the attack, what resources can the adversary use and end goal of the attack. Two main components are described as below:
  - **Attacker Goals:** Goals define what the adversarial input seeks from the attack. Various security goals such as integrity attack, availability attack, targeted attack, exploratory attack come under goals. An attacker can attempt a single goal or multiple at a given time.

  - **Attacker Capabilities:** Information at adversary's disposal like the training data (with or without labels), model parameters, number of queries adversary can make come under capabilities of the attacker. Also, training time attack and post-training attack are another aspect in which attacker's capabilities are defined.
- **White-box Attack:** In a white-box attack the adversary knows everything about the target model. The knowledge of adversary includes the learned weights, parameters used to tune the model. Labelled training data is also available in some cases. With this information the usual strategy of attacker is to model the distribution from weights and derive perturbed inputs to violate the boundaries while being in the *L2* limits.
- **Black-box Attack:** Contrast to white-box attacks a black-box attack has limited knowledge of the model, with no labelled data sometimes. An attack with black-box constraints is often modeled around querying the model on inputs, observing the labels or confidence scores.
- **Transferability:** It refers to the occasion of using a model trained on one the same of data, by observing inputs and outputs (querying) of the original model called "substitute model", and using the perturbed inputs from this model to attack the original "target model". Assumption here is that the substitute model simulates the target model.

1.2 Related Surveys, how our survey is different

In an early work [2], the authors introduced the now-standard taxonomy of Adversarial attacks in Machine Learning, the survey gave a structure to the interaction between attacker and defender. By fitting the existing adversarial attacks into this threat model, authors generalized on the principles involved in the security of Machine Learning. Another major contribution of the work is a line of defences to Adversarial attacks, that the authors presented based on their taxonomy. The work [3] also generalised and standardised the work done in this domain.

As the adversarial attacks became more frequent, later surveys in the field concentrated in specifics of domain. [4] is an extensive survey on attacks in Computer Vision with Deep Learning, [5] gave a data driven view of the same, and [6] studied the fundamental security policy violations in adversarial attacks.

As mentioned in Section 1, Black Box adversarial attacks are more close to real life and a lot of work is being carried out in this domain. However extensive and informative the surveys above have been, there is a lack of study exclusive to Black Box attacks. With this work we want to survey the existing bench mark Black Box attacks on different domains. We also address how each of the domains like Computer Vision, NLP differ in terms of robustness towards attack and the challenges in approaching these models. Table 1 shows a list of works we discuss in this paper, discussing the milestone contributions in the Black Box Adversarial space.

1.3 Why black box

While black-box attacks limit the capabilities of the attacker, they are more practical. In a general sense security attacks are carried out on full grown and deployed systems. Attacks with intent to circumvent a system, disable a system, compromise the integrity are seen in real world.

Consideration of two parties, adversary and challenger is possible in this paradigm. Challenger being the party that trains model and deploys and adversary being the attacker whose intention is to break the system for a predefined goal. Multiple capability configurations which reflect real world behavior are possible in this setting. As an example if a robber evading from the crime scene wants to fool the surveillance system into tagging a different car number, or a criminal targeting the prey's car number getting tagged are configuration changes to adversarial user goals. However, the system under attack in this example is owned by a third party. Neither the robber nor the criminal were part of the development of the surveillance system.

Month	Year	Method	Proposed Approach
Feb	2016	Papernot et al.[7]	Local substitute model
Dec	2016	Narodytska et al.[8]	Gradient-free
Jul	2017	Narodytska et al.[9]	Advanced Local Search
Nov	2017	Chen et al.[10]	ZOO
Feb	2018	Brendel et al.[11]	Boundary Attack
Jul	2018	Ilyas et al.[12]	Limited Queries and Info
Jul	2018	Cheng et al.[13]	Opt-Attack
Oct	2018	Bhagoji et al.[14]	Query Reduction using Finite Differences
Oct	2018	Du et al.[15]	Attack using Gray Images
Jan	2019	Tu et al.[16]	AutoZOOM
Apr	2019	Shi et al.[17]	Curls & Whey
Apr	2019	Dong et al.[18]	Translation-Invariant
May	2019	Chen et al.[19]	POBA-GA
May	2019	Brunner et al.[20]	Biased Sampling
May	2019	Li et al.[21]	NATTACK
May	2019	Moon et al.[22]	Combinatorial Optimization
May	2019	Ilyas et al.[23]	Bandits &Priors
Jul	2019	Alzantot et al.[24]	GenAttacks
Aug	2019	Guo et al.[25]	SimBA

Table 1. A comparison of works that have focused on black-box scenarios in an adversarial setting.

The limited capabilities of adversary are just a reflection of real world scenarios. Hence, works in this domain are more practical.

1.4 Evaluation of Attack

Adversarial attacks in general are evaluated based on the number of queries made on the model to converge on parameters. Lesser the number of queries better will be the attack, as time taken to start an attack will be minimized. Apart from time taken to craft an attack, perturbation norm is a standard measure of determining effectiveness of an attack. Borrowed from Machine Learning literature, the most common perturbation norms in user are l_2 and l_∞, which are standard error measures.

- l_2 **norm:** Also called as Euclidean norm, l_2 is the shortest distance between two vectors. In Adversarial setting, the distance from original input i and the perturbed input i‘ is calculated. With an aim to maintain the distance at minimum, the goal is to change an input i in a way indistinguishable by a human annotator.
- l_∞ **norm:** It is the highest entry in the vector space. l_∞ essentially determines the maximum value in the vector given.

2 BLACK BOX VS WHITE BOX

As discussed in Taxonomy 1.1, resources with an adversary differ considerably based on the type of attack model. In a White-box attack an adversary can have full access to the model right from the start of training, which leads to the availability of training data, testing data, network architecture, parameters, and finally the learned weights of a model etc. Also, the number of queries an adversary can make comes under resources. For example in a one-shot adversarial attack the attacker has

only a single shot at fooling the model, in such cases crafting the adversarial example does not involve a fine-tuning step.

When it comes to Black-box attacks the resources with adversary are considerably less. For one, the adversary will have no access to the model in training phase, nor the adversary knows weights and parameters used in the model. However, once the model is deployed, based on the type of attack the probability of each predicted label will be provided to the adversary. A stricter case is where only the predicted label is known to the attacker without any confidence score. Varying degree of information like labelled dataset are provided to adversaries in certain threat models. Again, as said above number of queries an attacker can make while crafting an adversarial example is considered under resources. An attack is considered superior if the amount of resources consumed is minimal.

3 CRAFTING A BLACK BOX ATTACK

A Black Box adversarial example starts with evaluating evaluating access to the resources at attacker's disposal. These include the kind of model, any model related parameters such as confidence scores, test set, training dataset etc. Once the resources are in place. An attack is carried out in one of the following ways. Using a **transferable** attack strategy, the adversary can chose to train a parallel model called a **substitute model**to emulate the original model. Attacker can use a much superior architecture than the original for the weight estimation. However, in the absence of a substitute model attacker chooses to have a **Query Feedback mechanism**, where attacker continuously crafts the perturbed input while querying on the model under attack.

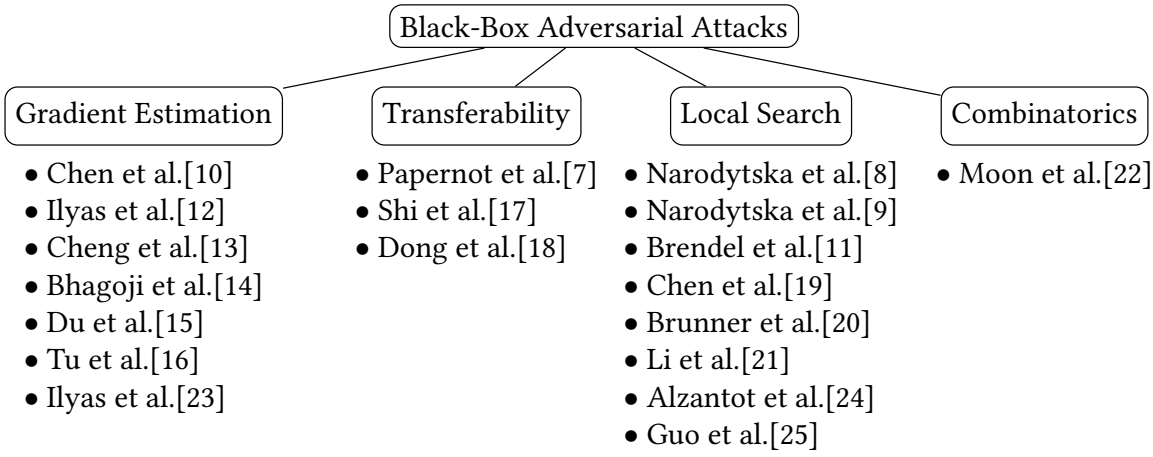
As mentioned earlier, when a network's architecture and model weights are unknown, which is the case with Black-box adversarial attacks, it is common practise for attackers to train another model from scratch, called a substitute model with a goal to emulate the model under attack. Once the substitute model with attacker has achieved satisfactory accuracy, by examining the weights adversarial examples are crafted. As the perturbed inputs start mis-classified by the model at attacker's hand, these inputs are used on actual model to achieve the adversarial goals. The choice for a substitute model is usually kept at a much superior level so as to understand the latent space with superior contrast. As the adversary has much control and knowledge of the task at hand, this is the choice for one-shot and targeted adversarial attacks. To put in a sentence: Transferring the learned weights by continuous querying, and strategically attacking the model comes under transferable attacks.

Adversarial attacks Without using a substitute model pose the challenge of identifying the boundaries, crafting the perturbed input at a manual level. Rather than training a model, then examining the weights this kind of attack is more hands on and simultaneous. The attack takes a query feedback form, where the attacker starts with a random input and starts adding noise to the input under the acceptable perturbation error level. As the confidence score of input deteriorates, attacker further pushes the noise in the direction of noise, more like following gradient. In some cases gradient descent itself is used to mark the direction and track the movement of noise in the right direction. Also, termed as local search the technique boils down to searching the right dimension in the latent space to get a mis-classified input. Detailed discussion of attacks following both the substitute model and query feedback mechanism have been the pillars of Black Box adversarial attacks. Section to follow discuss the same in a great detail, Table **??** shows the works based on domain that we cover in this paper.

4 BLACK BOX ATTACKS IN COMPUTER VISION

This section provides a detailed discussion of recent works, and important works in the Black Box adversarial attacks domain. Subsections are organized in the order of early works appearing

Classification of prior work on black-box scenario-based attacks.



first. In these sections we cover the dataset used, key contributions, loss function used to calculate perturbations are discussed.

4.0.1 Popular Image data sets. The most frequently used datasets to demonstrate Adversarial attacks are as follows:

- **MNIST**: A popular image dataset of black and white, handwritten digits. It has 10 classes of digits from 0 to 9. This is considered as an entry level dataset.
- **CIFAR 10**: The dataset consists of 60,000 color images in 10 classes. This dataset is large and diverse in style compared to MNIST.
- **ImageNet**: With 14 million images across 27 high level categories and 21,841 sub level categories. This dataset played an important in computer vision, and it is an advanced dataset.

4.1 Attack Techniques

We divide the proposed attack methodologies into 4 categories as shown in **??**.

4.1.1 Gradient Estimation.

- **ZOO**: Chen et al.[10] proposed *Zeroth Order Optimization(ZOO)* to estimate the gradients of target DNN in order to produce an adversarial image. The threat model assumed by the authors is that the target model can only be queried to obtain the probability scores of all the classes. The loss function formulated is:

$$ForTargetedAttack : f(x,t) = max\{max_{i\neq t} log[F(x)]_i - log[F(x)]_t, -k\} \quad (1)$$

$$ForUntargetedAttack : f(x) = \{log[F(x)]_{t_0} - max_{i\neq t_0} log[F(x)]_i, -k\} \quad (2)$$

The author then uses *symmetric difference quotient*(cite) to estimate the gradient $\frac{\partial f(x)}{\partial x_i}$:

$$\hat{g} := \frac{\partial f(x)}{\partial x_i} \approx \frac{f(x+he_i) - f(x-he_i)}{2h} \quad (3)$$

The above naive solution requires querying the model $2p$ times where p is the dimension of the input. So the author proposes 2 *stochastic coordinate methods*: ZOO-Adam and ZOO-Newton in which a gradient is estimated for a random coordinate and the update formula is obtained using ADAM(cite) and Newton's Method(cite) until it reaches convergence. The

authors also discuss about generation of noise in a lower dimension in order to improve efficiency and specify its advantages and disadvantages.

- **Limited Queries & Information**: The authors in [12] take three primary cases into account to devise successful black-box adversarial attacks. The first case talks about the constraint of limited queries that can be made to the model by the adversary. This could again be of two types- one that contains a time limit and the other that is concerned with the monetary limit. The authors present a variant of Natural Evolution Strategies(NES) coupled with Projected Gradient Descent (PGD), as used in white-box attacks, to construct adversarial examples.
  The second case is modelled around the constraint of partial-information setting where the adversary can obtain confidence scores of the first k classes for a given input. These scores may not add up to 1 since the adversary doesn't have access to the probabilities for each possible classification label. To tackle this scenario, instead of beginning with the input image x, it is recommended to begin with x_o belonging to the target adversarial class y_{adv}. Hence, after each step, we need to ensure two things-
  – The target adversarial class needs to remain in the top-k classes at all points in time when the input image is being perturbed.

$$\epsilon_t = min\, \epsilon' \; s.t. \; rank\left(y_{adv} | \prod_{\epsilon'}(x^{(t-1)})\right) < k \tag{4}$$

  – The probability of the input image getting classified as the target class increases with every iteration of PGD.

$$x^{(t)} = arg\max_{x'} P\big(y_{adv} | \prod_{\epsilon_{t-1}}(x')\big) \tag{5}$$

  Lastly, in the third scenario the adversary is not given any confidence scores. Rather, the adversary can only obtain the names of the classification labels for the given input data. The authors define a discretised score $R(x^t)$ for an adversarial example to quantitatively represent the adversarial nature of the input image at each step t, with access only to the $top-k$ sorted labels.
  The proposed approach was evaluated on InceptionV3 network with 78% $top-1$ accuracy and also on Google Cloud Vision (GCV) API which presents a real-world scenario to perform an adversarial attack. 90% and above accuracy is achieved for all three settings when the InceptionV3 network is attacked in the above-mentioned techniques.

- **Opt-Attack**: Cheng et al.[13] attempt to devise a black box adversarial attack in a much stricter hard-label setting. This means that querying the target model gives only the target label unlike other threat models where K-class probability scores or labels are considered. The authors make the attack query-efficient by treating the problem as a real valued, continuous optimization problem.
  The objective function proposed by them is as follows:

$$UntargetedAttack: \quad g(\theta) = min_{\lambda>0}\, \lambda \;\; s.t \;\; f(x_0 + \lambda * \frac{\theta}{||\theta||}) \neq y_0 \tag{6}$$

$$TargetedAttack(giventargett): \quad g(\theta) = min_{\lambda>0}\, \lambda \;\; s.t \;\; f(x_0 + \lambda * \frac{\theta}{||\theta||}) = t \tag{7}$$

  Here θ is the search direction and $g(\theta)$ is the distance from the input image x_0 to the closest adversarial example in direction θ. Chen et. al. uses $\theta^* = argmin_\theta\, g(\theta)$ to obtain $x^* = x_0 + g(\theta^*) * \frac{\theta}{||\theta||}$ by using a coarse grained search to initially find a decision boundary and

then fine tuning the solution using Binary Search. The optimization problem proposed above is solved using *Randomized Gradient Free (RGF) Method* which is an iterative Zeroth Order Optimization(ZOO) Method in which the small increment value $\boldsymbol{u}_t$ is sampled randomly from a Gaussian Distribution.

- **Query Redction using Finite Differences**: Finite Differences(cite) was used to estimate the gradient in the case of black-box attacks where no knowledge regarding the gradient of the loss function exists. The authors in [14] propose a two-sided gradient estimation for a function $f(x)$, where $x \in R^d$.
  The loss function used in this scenario is based on logit, which was proved to work well in the white-box attacks [26]. The logit loss is given by:

$$l(x, y) = \phi(x + \delta)_y \; - \; max \; \{\phi(x + \delta)_i \; : i \neq y\} \tag{8}$$

  where y represents the actual label for a given sample x and $\phi(.)$ are the logits. By taking the logarithm of the SoftMax probabilities, the logit values can be computed up to an additive constant. This additive constant gets cancelled out because the difference in the logits is equal to the loss function. Hence, the proposed method of using Finite Differences can be used to calculate the difference between the logit values of the actual label y and the next most likely label y'.
  A drawback to this naïve approach is that the model is required to make $O(d)$ queries per input, where d is the dimensions of input. Authors have further proposed two query reduction techniques based on grouping the features. First approach is based on random selection of features and grouping them together, which will allow for simultaneous gradient estimation for the selected features. Second approach is more technical, since it involves computing directional derivatives along the principal components as determined by principal component analysis (PCA).
  The proposed approach was evaluated on two data sets - MNIST and CIFAR-10. Two models, one with 2 convolutional layers attached to a fully connected layer and, another with 3 convolutional layers. These models were trained on the MNIST dataset. On the other hand, ResNet-32 and ResNet-28-10 models were used for training in the case of CIFAR-10 data set. The conclusion obtained from this approach is that the effectiveness of the crafted adversarial examples can be reduced by modifying the model output probabilities on which Gradient Estimation attacks are hugely dependent on.

- **Attack using Gray Images**: A completely new approach was followed in [15] to generate black-box adversarial examples by initializing the adversarial sample from a gray image. This is followed by the fact that the adversary doesn't have access to any natural input. This methodology begins by initializing the input for the target model with normalized pixel values within [0,1]. Hence, the adversary now has more flexibility in adding perturbation to the image by either increasing the pixel value to 1 to make it brighter or decreasing the pixel value to 0 to make it darker.
  Then the authors have defined the fitness function in this approach as:

$$J(\theta) = P(y'|\theta) \approx [F(\theta)]_{y'} \tag{9}$$

  The objective here is to maximize the probability term given in the fitness function until the input image gets classified into another class y'. Similar to many other approaches, the authors have preferred to adopt NES optimization. Here θ represents the perturbed input

and Gaussian search distribution has been chosen given the fact that it provides a simpler method of estimating the gradient where the two parameters that need to be optimized are mean ϕ_d and variance σ_{dxd}.
An important point to be noted in this approach is that this is a region-based attack algorithm which implies that the gradients are dependent on the region space. Evaluation of gradients is then followed by performing a projected gradient ascent similar to the approach in [12, 27]. Input update can be shown as:

$$x^t = \prod_{[0,1]}(x^{t-1} + \lambda sign(g_t(h, w))) \tag{10}$$

where gamma represents the learning rate and $\pi_{[0,1]}$ represents the projection operator which clips the values of the pixel in the range [0,1]. The search gradients are used to update the image iteratively until the goal is achieved.

- **AutoZOOM**: Tu et al.[16] proposed the framework AutoZOOM, which is short for Autoencoder based Zeroth Order Optimization,for crafting efficient Black Box Attacks. The two main contributions of this approach were: (i) an adaptive random gradient estimation method for stabilizing query counts and distortion and (ii) an autoencoder trained offline on unlabelled data or a simple Bilinear Image Resizer(BiLIN) in order to expedite the attack process.
The optimization problem formulated by Tu et al.[16]:

$$min_{x\in[0,1]^d} Dist(x, x_0) + \lambda.Loss(x, M(F(x)), t) \tag{11}$$

where $Dist(x, x_0) = ||x - x_0||_p$, $F(x) : [0, 1]^d \rightarrow \mathcal{R}^K$, $M(.)$ is the monotonic function applied on classifier outputs, $Loss(.)$ is the attack objective reflecting the likelihood of predicting $t = argmax_{k\in\{1,2,\dots K\}}[M(F(x))]_k$ and λ is the regularization constant. Loss function can be both either training loss of DNN or C&W Loss[28].
Inspired by the convergence of *Symmetric Difference Quotient Method*[29] and taking query efficiency into consideration, the authors proposed a scaled random full gradient estimator of $\nabla f(x)$, defined as:

$$\mathbf{g} = b.\frac{f(x + \beta u) - f(x)}{\beta}.u \tag{12}$$

To control error in gradient estimation, an average is taken over q random directions:

$$\bar{\mathbf{g}} = \frac{1}{q}\sum_{j=1}^{q}\mathbf{g}_j \tag{13}$$

They have also proven that l_2 distance between $\bar{g}$ and $\nabla f(x)$ is constrained by an upper bound:

$$\mathcal{E}||\bar{g} - \nabla f(x)||_2^2 \leq 4(\frac{b^2}{d^2} + \frac{b^2}{dq} + \frac{(b-d)^2}{d^2})||\nabla f(x)||_2^2 + \frac{2q+1}{q}b^2\beta^2L^2 \tag{14}$$

The above upper bound is minimum when $b \approx q$. They also mention that the convergence rate of zero order gradient descent methods is $O(\sqrt{\frac{d}{T}})$[30][31][32] and so it is plausible to generate perturbation from a lower dimension. In AutoZOOM, random gradient estimation is performed from a lower dimension $d' < d$ in order to improve query efficiency. A decoder,$D : \mathcal{R}^{d'} \rightarrow \mathcal{R}^d$ is used to add noise such that: $x = x_0 + D(\delta')$ where $\delta' \in \mathcal{R}^{d'}$. Initially $q = 1$ but with each iteration q is increased in order to get more accurate estimation.

- **Bandits & Priors**: Ilyas et al.[23] successfully exploited prior information about the gradient using bandit optimization. The challenge of obtaining prior information about the gradient can be met by two key observations :-
  – The input data point for which gradient computation is carried out is not arbitrary and it's structure is predictable and also reflected in the gradient.
  – Heavy correlation can be found among the gradients that are used in successive iterations while performing iterative gradient attacks (e.g. PGD).

  Two categories of priors, namely- time-dependent priors and data-dependent priors have been introduced which significantly helped in performing improved black-box adversarial attacks. The value of cosine similarity at each optimization step between successive gradients proves the correlation between them. Hence, the authors used the gradient at time $t-1$ as a prior for the gradient calculation at time t.

$$\frac{\left\langle \nabla_x L(x_t, y), \nabla_x L(x_{t+1}, y) \right\rangle}{\left\|\nabla_x L(x_t, y)\right\|_2 \left\|\nabla_x L(x_{t+1}, y)\right\|_2} \quad t \in \{1 \ldots T-1\} \tag{15}$$

  On the other hand, an analogy to the correlation between pixel values of an image has been made to understand the correlation between gradients for two coordinates that are close. Hence, the structure of the input data point can also be used to reduce the query complexity. Untargeted adversarial examples have been crafted by considering both l_2 and l_∞ norm-based threat models on the ImageNet data set. The approach was experimented on three classifiers-Inception-v3, ResNet-50 and VGG16. An average of 1117 queries were required in the case of the proposed approach that made use of bandit optimization along with time and data-dependent priors as compared to about 1735 queries in the simple case of NES. Attack failure rate was decreased from 22.2% in l_2 and 34% in l_∞ for NES to 4.6% in l_2 and 15.5% in L_∞ for the Bandit approach.

4.1.2 *Transferability.*

- **Local Subsitute Model**: In [7], the targeted deep learning model O, also known as the oracle was used to construct a synthetic data set. Labels for this synthetically crafted input data were obtained by querying the oracle O itself. This assists the attacker to create an approximation F of the target model O. The concept of transferability allows the attacker to ensure that the input images misclassified by the approximated model F, will also be misclassified by O. The primary challenge faced by the attacker in this approach is,to train the model F without any prior knowledge of the architecture O. Constraint on the number of times the attacker can query the model O was a limitation as well. To overcome this problem, the authors propose Jacobian-based Dataset Augmentation(add ref) which helps in approximating the decision boundaries with limited queries.

  Authors proposed a five step approach, as listed below:

(1) The attacker collects a small set S_o which represents the entire set of inputs S to the oracle. For example, a limited number of input images, say k from every class will be obtained.
(2) Then a suitable architecture is decided upon that can be used to train the approximation model F. It is shown that the number, size and type of layers do not determine a successful attack.
(3) The substitute model is trained by repetitively applying the following steps-
  (a) Model F queries the oracle model O for each of the samples $\vec{x} \in S_p$, where S_p is the training set available to the substitute the model.
  (b) Approximated model is then trained by the attacker using the available training set S_p.

(c) Apply the augmentation technique on the training set S_p available with the model F and use it to generate new synthetic data points. This results into an even larger training data set S_{p+1} which is a better representative of the decision boundaries of the target model.

The augmentation technique proposed, i.e. the Jacobian-based Dataset Augmentation is formulated as:

$$S_{p+1} = \{\vec{x} + \lambda \cdot sgn(J_F[\check{O}(\vec{x})] : \vec{x} \in S_p)\} \cup S_p \tag{16}$$

where, λ is the step-size taken during augmentation to create the set S_{P+1} from the initial available training set S_p.

- **Curls & Whey**: The authors in[17] aim to enhance the diversity as well as transferability of adversarial examples in the case of black-box attacks. Their approach consists of two steps. The first step proposed deals with the problem of adding noises/perturbation to an input sample monotonically along the gradient ascent's direction using Curl's iteration, whereas the second step deals with the issue of removing excessive redundant noises from the crafted adversarial examples using Whey optimization. The proposed methodology can be better understood through [17].
  The aim of the methodology is to push the input data point x from the current class to a different class with which it shares its decision boundary. To perform the same operation, noise may be added to this data point monotonically along the direction of gradient ascent. However, the authors propose a shortcut to reach a nearer point that is present across the decision boundary. They suggest to initially follow the direction of the gradient descent and reach the 'valley floor' and then perform iterations to climb up the gradient. This is known as the Curl's iteration. Noise refinement mechanisms are followed after and before the Curl's iteration respectively, which has been defined as Whey optimization.

- **Translation-Invariant**: This attack[18] method makes use of an ensemble of images rather than aiming to optimize the objective function. The calculation of the gradient makes use of the assumption that convolutional neural networks possess the translation-invariant property (cite). Hence, the authors propose to shift the original input image by a limited number of pixels (10, in this case) in a 2-dimensional space.
  The proposed method can be integrated with any of the gradient-based attacks which includes Fast Gradient Sign Method (FGSM) (cite), Basic Iterative Method (BIM) (cite), Momentum Iterative Fast Gradient Sign Method (MI-FGSM) (cite), Diverse Inputs Method (cite) and Carlini & Wagner's Method (C&W) (cite). Experiments have been shown by the integration of the translation-invariant property with the FGSM attack where the update rule can be stated as:

$$x^{adv} = x^{real} + \epsilon \dot{s}ign(W * \nabla_x J(x^{real}, y)) \tag{17}$$

where, the notations have the usual meanings as described in (cite FGSM paper).

4.1.3 *Local Search.*

- **Gradient-free** Authors in [8] propose an adversarial attack methodology based on the concept of greedy local-search technique. Their proposed algorithm perturbs an image by choosing the most relevant pixels of the image to which noise is added. The observed change by adding this noise in each iteration is used to approximate the loss function's gradient. The importance of the pixels can be understood in this step by observing the change in classification accuracy after each round.

The most important contribution of this work is the creation of the neighbouring images to a given input image. This neighbourhood will consist of all images that are different from the previous round's image by one pixel only. Initially, a pixel location is randomly selected and perturbation is added to it. The next pixel location is chosen among all the pixels that lie within a square whose side length is $2p$ and centre is the previous pixel location, where p is an arbitrary parameter.
The authors have focused on $k - misclassification$ where the correct label for an image should not be enlisted in the top-k classes predicted by the model. As soon as the target model is able to push the correct class out of the top-k classes, no more noise is added to the image and the algorithm gets terminated. The local search technique adopted is modified to increase the classification score of the targeted class in each round, in the case of targeted attacks.

- **Advanced Local Search**: Local search technique is a commonly used incomplete search algorithm used to solve problems pertaining to combinatorics. In [9], the authors have used this strategy over the image space to construct an approximated network gradient. The objective function is to minimize the probability that the perturbed image does not share the classification label with the original image. The cost function defined is-

$$f_{c(I)}(\hat{I}) = o_{c(I)} \tag{18}$$

where $NN(\hat{I}) = (o_1, ..., o_C)$, in which NN is the target neural network model and o_j represents the probability of image $\hat{I}$ belonging to class j. This function is minimized by the local search technique.
The next step in the formulation is to create a set of images that can act as neighbours of the original image. The authors have defined these neighbour images with the help of pixel locations since images generated after a single iteration differ only in one pixel from the previous image. Consider the pixels that were perturbed in the previous round. Hence, in the current round, all those pixels that lie in a square of side length $2d$ with the perturbed pixel at its centre will be considered. Here, d is only a parameter.
In the last step, a function g has been defined as the transformation function which generates a new image $\hat{I}_i$ after every round i by adding perturbation to a set of pixels only.

- **Boundary Attack**: Decision-based[11] adversarial attacks that are based on a model's final decision, i.e. class labels or transcribed sentences, are more common in real-life scenarios as compared to confidence score-based attacks. The authors also argue that these attacks are more robust to defense techniques such as gradient masking, intrinsic stochasticity or robust training. Also, lesser information regarding the targeted model is required to conduct a successful attack. This was the first time that decision-based attacks were introduced that focus on deep learning models using real-life data sets such as ImageNet.
The proposed algorithm initialises by taking a random distribution of pixels for an image, in the range of [0, 255]. Using this image, the algorithm traverses the decision boundary of the class to which the sample input image belongs to. The initialised image is then able to reduce its distance w.r.t. the input image besides staying inside the adversarial region.

- **POBA-GA**: Chen et al.[19] proposed three parameters to be used to generate different perturbations- different noise point pixel thresholds, number of noise points, and noise point size. The next step is to add this noise to the initial adversarial example to generate the next

generation of population. For evaluation purpose, a fitness function ϕ is defined followed by checking if the adversary has satisfied the termination condition or not. Lastly, to generate the next set of population for optimizing the perturbation, the suggested approach follows the typical operations of a genetic algorithm, i.e. selection, crossover and mutation.
The importance of the proposed fitness function is that it helps to evaluate how effective the adversarial input is at a given iteration. Hence, the convergence of the model largely depends on this function. The authors have incorporated two important qualities of adversarial images into the definition of the proposed function which are

$$\phi(AS_i^t) = P(AS_i^t) - \frac{\alpha}{maxZ(A^0)} Z(A_i^t) \tag{19}$$

where the left hand side of the equation represents the fitness function for Adversarial Sample i, (AS_i) is the probability term which accounts for the effectiveness of the attack for the sample to get misclassified and $Z(A_i^t)$ is a metric originally proposed in this methodology to account for the perturbation introduced. α is used as a parameter to control the proportion for the fitness function between attack performance and the perturbation size. For the purpose of reducing the number of queries to the target model, alpha is initially taken as zero and later assigned a value when the attack is successful. The fitness function finally looks like-

$$\phi(AS_i^t) = \begin{cases} p(y_1|AS_i^t) - p(y_0|AS_i^t) - \frac{\alpha}{maxZ(A^{t_0})} Z(A_i^t) & y_1 \neq y_0 \\ p(y_2|AS_i^t) - p(y_0|AS_i^t) & y_1 = y_0 \end{cases} \tag{20}$$

- **Biased Sampling**: The authors in [20] adopt a unique approach to introduce perturbations for creating adversarial samples known as Perlin Noise[33]. This method was originally used to generate procedural texture patterns for computer graphics but can also be used to create parametrized patterns having low-frequency. In the case of an original Boundary Attack[11], an orthogonal perturbation η^k is applied along a hyper-spherical surface present around the original image. However, Perlin noise distribution is used to create a sample to introduce a low-frequency prior into the Boundary Attack. Usually, a permutation vector v with a size of 256 is used to parametrize the noise. The noise patterns are sampled in two directions- along the height and width of the image due to which the patters get strongly concentrated in the directions of low-frequency.
Image masking has been performed to calculate the difference between the pixel values of the original image and the adversarial image and can be shown as:

$$M = \left| X_{adv} - X_{orig} \right| \tag{21}$$

This calculation is performed after every step and applied on every single element of the orthogonal perturbation η^k that was sampled previously.

$$\eta^k_{biased} = M \odot \eta^k \; ; \; \eta^k_{biased} = \frac{\eta^k_{biased}}{\left\| \eta^k_{biased} \right\|} \tag{22}$$

Hence, distortion of only the pixels with a large difference value is increased and it remains the same for the remaining pixels further reducing the search space effectively.

- **$\mathcal{N}$-Attack**: A black box adversarial attack has been proposed in [21] which tries to learn the probability density over a small region (l_p *ball*) centered around the input such that a sample

drawn from the density is likely an adversarial image. Their work is similar to Ilyas et al.[12] who have employed the loss function: $f(x') := -F(x')_y$. They have argued that Ilyas et al's [12] algorithm is inferior as it relies on accurate estimation of gradients which is not possible when DNN is non-smooth. For estimating gradients $\nabla f(x_t)$ with derivative-free methods the authors have also used *NES*[34].
The smooth optimization criteria proposed by the authors is:

$$min_\theta J(\theta) := \int f(x')\pi_S(x'|\theta)dx' \tag{23}$$

where $f(x')$ is the C&W Loss[28], $\pi_S(x'|\theta)$ is the probability density in region S (l_p *ball*) i.e $||x' - x||_p \leq \epsilon$ and also $dim(\theta) << dim(x)$. The authors reformulated the problem with change of variables:

$$x' = proj_S(g(z)), \quad z \sim \mathcal{N}(z|\mu, \sigma^2) \tag{24}$$

where $g : \mathbb{R}^{dim(\mu)} \mapsto \mathbb{R}^{dim(x)}$. The transformation steps involve sampling z i.e $z \sim \mathcal{N}(\mu, \sigma^2)$, then $g(z)$ is computed: $g(z) = \frac{1}{2(tanh(g_0(z))+1)}$. After this, $\delta' = clip_p(g(z) - x)$ for p = 2 or ∞:

$$clip_2(\delta) = \begin{cases} \delta\tau_2/||\delta||_2 & if||\delta||_2 > \tau_2 \\ \delta & else \end{cases} \tag{25}$$

$$clip_\infty = min(\delta, \tau_\infty) \tag{26}$$

Updated $J(\theta)$ due to variable transformation:

$$J(\theta) = \mathbb{E}_{\mathcal{N}(z|\mu,\sigma^2)} f(proj(g(z)))$$

$\theta = (\mu, \sigma^2)$ are the unknowns. σ is found using Grid Search and *NES*[34] is used to find μ:

$$\mu_{t+1} \leftarrow \mu_t - \eta \nabla_\mu J(\theta)|_{\mu_t}$$

$$\mu_{t+1} \leftarrow \mu_t - \frac{\eta}{b}\sum_{i=1}^{b} f(proj_S(g(z_i)))\nabla_\mu log\mathcal{N}(z_i|\mu_t, \sigma^2) \tag{27}$$

The process is iteratively repeated until arriving at an sampled image x' such that $C(x') \neq C(x)$.

- **GenAttack**: Alzantot et al.[24] uses genetic algorithms, which are population based gradient-free optimization strategies in order to produce an adversarial image. The author justifies using the above approach by mentioning the overhead produced by excessive querying in *Gradient Estimation Methods* and also the added advantage of being able to bypass defenses based on Gradient Masking and Obfuscation. Genetic Algorithm is based on the process of natural selection which involves iteratively choosing the $\boldsymbol{P}$ set of candidates which constitute a *generation* in each iteration. The quality of the candidates are assessed using a *fitness* function and the candidates for the next generation are obtained using *crossover* i.e breeding of parents selected randomly to produce a child and *mutation* which is the noise added to the child in order to reflect diversity of population. The fitness function proposed here can be shown as:

$$Fitness(x) = logf(x)_t - log\sum_{j=0, j\neq t}^{j=k} f(x)_c \tag{28}$$

Here $f : \mathbb{R}^d -> [0,1]^K$ is the black box classifier which outputs probabilities of the K classes and t is the label for targeted attack. The parents for crossover are randomly sampled using

probs, i.e. the probabilities which are obtained using softmax of the fitness values. Crossover operation involves selecting features of $parent_1$ and $parent_2$ with $(p, 1-p)$ respectively where $p = \frac{fitness(parent_1)}{fitness(parent_1)+fitness(parent_2)}$. A Child is mutated using the following formula $child_{mut} = child + Bernoulli(p) * \mathbf{U}(-\alpha\delta_{max}, \alpha\delta_{max})$. α, ρ and δ_{max} are mutation range, mutation probability and l_∞ metric respectively.

- **SimBA**: The study shown in [25] proposes a very simple yet efficient approach to generate adversarial examples in a black-box setting. From a set Q of orthonormal vectors, the algorithm picks any single vector q and the modification made to the input image x can be expressed as $x + q\epsilon$ or $x - q\epsilon$. The idea behind this is to modify and add noise to the image in a particular direction and check if this brings about a change in the probability $P(y|x)$ which represents the probability of the label being y given the input image x. Here, ϵ represents the step size and can be controlled by the adversary. To keep the number of queries to the target model minimum, the methodology proposed ensures that when the algorithm picks $q \in Q$, there are no two directions that cancel out each other.
  Inspired by the work of Guo[35], the authors suggest to make use of low frequency noise to be added to the input image for generating better adversarial examples. They employ the discrete cosine transform(DCT) to map signals in a 2D image space R^{dxd} to frequency coefficients belongs to cosine wave function magnitudes.

4.1.4 *Combinatorics.*

- **Combinatorial Optimization**: In the approach presented in [22], the optimization problem has been treated as a discrete surrogate problem. The authors support this my mentioning that first order attacks like FGSM[36] and PGD[37] have a good performance due to their constrained nature and also due to querying the gradient of the loss with respect to the input. Using the Taylor Expansion the loss and the optimization problem becomes:

$$l(x_{adv}, y) \approx l(x, y) + (x_{adv} - x)^T \nabla_x l(x, y)$$

$$maximize_{||x_{adv}-x||_\infty \leq \epsilon} \Rightarrow maximize_{x_{adv}} {x_{adv}}^T \nabla_x l(x, y) \quad subject\ to\ -\epsilon\mathbf{1} \preceq x_{adv} - x \preceq \epsilon\mathbf{1} \quad (29)$$

Due to the bounds present in the equation 29, the authors concluded that the solution of the above linear program would be obtained at the vertex of the l_∞ ball. So the reformulated optimization problem becomes:

$$maximize_{S \subseteq V} \left\{ F(S) \triangleq f\left(x + \epsilon \sum_{i \in S} e_i - \epsilon \sum_{i \notin S} e_i \right) \right\} \quad (30)$$

where $f(x) = l(x, y_{gt})$ for untargeted attacks and $f(x) = -l(x, y_{target})$ for targeted attacks. They then explain the property of submodularity, a function $F : 2^V \rightarrow \mathcal{R}$ is submodular if for every $A \subseteq B \subseteq V$ and $e \in V \setminus B$: $\Delta(e|A) \geq \Delta(e|B)$. So submodular functions exhibit a diminishing property as set size increases. With submodularity and using greedy style algorithms, it is possible to obtain $(1 - \frac{1}{e})$ approximation of global optimum for monotone submodular functions and $\frac{1}{3}$ approximation for non-monotone submodular functions. The authors use a modified version of local search algorithm by Fiege et al.[38] which alternates between greedily inserting the element while the marginal gain is strictly positive $\left(\Delta\left(e|S > 0\right)\right)$ and removing the element while the marginal gain is also strictly positive.

4.2 Comparative Analysis of Black Box attack techniques

The attack methods proposed have been tested quantitatively on three primary datasets as mentioned earlier, i.e. MNIST, CIFAR-10 and ImageNet. Therefore, we present a qualitative comparison between the results of these techniques based. The results can be first divided broadly into two categories: Targeted & Untargeted attacks for each of the three datasets. In each table, we mention the model used for generating the adversarial attack, average number of queries made to the targeted model, average L_2 distance between original image and the perturbed image, attack success rate and lastly some specific parameters mentioned in the experiments of the respective study. The results for different comparative parameters are obtained from common performance metrics as produced by the respective research studies.

Method	Model	Average # Queries	Average L_2	Attack Success Rate	Specific Parameters Used
Chen et al.[10]	InceptionV3	-	1.19916	88.90%	Step Size = 0.002, # Iterations = 1500
Brendel et al.[11]	ResNet-50	192102	4.876	100%	-
Cheng et al.[13]	ResNet-50	145176	4.934	100%	-
Moon et al.[22]	InceptionV3	722	-	98.50%	L_∞ setting, ϵ = 0.05, Max # Queries = 10,000
Guo et al.[25]	InceptionV3	1283	3.06	97.80%	Step Size = 0.2, # Iterations = 10,000

Table 2. Comparison of methods used for generating an untargeted attack on the ImageNet data set.

From Table 2, we can see that the attack success rate for [11] and [13] is 100%. However, the technique proposed by [22] takes the least number of queries to produce a successful black-box attack.

Method	Model	Average # Queries	Average L_2	Attack Success Rate	Specific Parameters Used
Chen et al.[10]	InceptionV3	-	3.425	-	Batch Size = 128, # Iterations = 20,000
Ilyas et al.[12]	InceptionV3	11550	-	99.20%	L_∞ setting, ϵ = 0.05
Du et al.[15]	InceptionV3	1701	24.323	100%	Image size-64x64, Momentum factor = 0.4
Tu et al.[16]	InceptionV3	13525	0.000136	100%	Regularization coefficient = 10
Chen et al.[19]	InceptionV3	3573	4.559451	95%	# Iterations = 400
	ResNet-50	3614	3.0575142	98%	# Iterations = 400
	VGG19	3786	4.023045	96%	# Iterations = 400
Brunner et al.[20]	InceptionV3	-	-	85%	Max # Queries = 15,000, L_2 setting, ϵ = 25.89
Moon et al.[22]	InceptionV3	7485	-	99.90%	L_∞ setting, ϵ = 0.05, Max # Queries = 100,000
Ilyas et al.[23]	InceptionV3	1858	-	84.50%	Max # Queries = 10,000, L_2 setting
	InceptionV3	1117	-	95.40%	Max # Queries = 10,000, L_∞ setting
	ResNet-50	993	-	90.30%	Max # Queries = 10,000, L_2 setting
	ResNet-50	722	-	96.60%	Max # Queries = 10,000, L_∞ setting
	VGG16	594	-	82.80%	Max # Queries = 10,000, L_2 setting
	VGG16	370	-	91.60%	Max # Queries = 10,000, L_∞ setting
Alzantot et al.[24]	InceptionV3	11081	61.68669	100%	Max # Queries = 1,000,000, L_∞ setting, ϵ = 0.05
Guo et al.[25]	InceptionV3	7899	9.53	100%	Step Size = 0.2, # Iterations = 30,000

Table 3. Comparison of methods used for generating a targeted attack on the ImageNet data set.

A lot of models such as Inception-V3, ResNet-50, VGG16 and VGG19 have been used to experiment targeted attacks on ImageNet data set and this can be seen in Table 3. The attacks mentioned in [15] and [24] suffer greatly in terms of average L_2 distance parameter as compared to the others. The authors in [23] used three different models and experimented their attack by keeping a limit on the max number of queries at 10,000 in L_2 and L_∞ settings. The number of queries were less than 2000 and in most cases below 1000 whereas the attack success rate ranged from 82.80% using VGG16 model to 96.60% in ResNet-50.

In Table 4, three out of four methods achieve a 100% success rate whereas only the attack proposed by [13] surpasses the others in the remaining two parameters.

Almost all the methods in Table 5 show a 100% attack success rate. The attacks proposed by [16], [19] AND [24] achieve an attack in less than 1000 queries.

Method	Model	Average # Queries	Average L_2	Attack Success Rate	Specific Parameters Used
Chen et al.[10]	-	-	1.4955	100%	Batch Size = 128, Step Size = 0.01, # Iterations = 3000
Brendel et al.[11]	ResNet-50	131972	1.11545	100%	-
Cheng et al.[13]	ResNet-50	67036	1.0827	100%	-
Bhagoji et al.[14]	-	-	2.7	99.7	Batch Size = 100

Table 4. Comparison of methods used for generating an untargeted attack on the MNIST data set.

Method	Model	Average # Queries	Average L_2	Attack Success Rate	Specific Parameters Used
Chen et al.[10]	-	-	1.9871	98.90%	Batch Size = 128, Step Size = 0.01, # Iterations = 3000
Brendel et al.[11]	ResNet-50	93543	2.0626	100%	-
Cheng et al.[13]	ResNet-50	59094	1.7793	100%	-
Bhagoji et al.[14]	Self-designed, accuracy = 99.2%	62720	-	100%	Batch Size = 100, L_∞ setting, ϵ = 0.3
Tu et al.[16]	C&W	2428	3.55936	100%	Regularization coefficient = 0.1
	C&W	730	3.23792	100%	Regularization coefficient = 1
	C&W	510	3.66128	100%	Regularization coefficient = 10
Chen et al.[19]	-	423	2.352	100%	# Iterations = 100
Alzantot et al.[24]	-	996	-	100%	Step Size = 1.0, Max # Queries = 100,000

Table 5. Comparison of methods used for generating a targeted attack on the MNIST data set.

Method	Model	Average # Queries	Average L_2	Attack Success Rate	Specific Parameters Used
Chen et al.[10]	-	-	0.19973	100%	Batch Size = 128, Step Size = 0.01, # Iterations = 1000
Brendel et al.[11]	ResNet-50	384204	4.8758	-	-
Cheng et al.[13]	ResNet-50	145176	4.9339	-	-
Bhagoji et al.[14]	ResNet-32	-	-	100%	Batch Size = 100
Moon et al.[22]	ResNet-32	1261	-	48%	Max # Queries = 20,000

Table 6. Comparison of methods used for generating an untargeted attack on the CIFAR-10 data set.

In Table 6, the black-box attack proposed by [10] achieves an exceptional performance by getting a 100% success rate while getting an average L_2 distance of 0.19973. [22] uses a ResNet-32 model to achieve an average of 1261 queries, however, the attack success rate is as low as 48%.

Method	Model	Average # Queries	Average L_2	Attack Success Rate	Specific Parameters Used
Chen et al.[10]	same as 14	-	0.39879	96.80%	Batch Size = 128, Step Size = 0.01, # Iterations = 1000
Brendel et al.[11]	ResNet-50	170972	0.2395	-	-
Cheng et al.[13]	ResNet-50	130020	0.2476	-	-
Bhagoji et al.[14]	ResNet-32	61440	-	100%	Batch Size = 100, L_∞ setting, ϵ = 8
Tu et al.[16]	C&W	1524	3.6864	100%	Regularization coefficient = 0.1
	C&W	332	0.00101	100%	Regularization coefficient = 1
	C&W	259	0.00115	100%	Regularization coefficient = 10
Chen et al.[19]	same as 4	381	0.2088	100%	# Iterations = 100
Alzantot et al.[24]		804	-	96.50%	Step Size = 1.0, Max # Queries = 100,000

Table 7. Comparison of methods used for generating a targeted attack on the CIFAR-10 data set.

As shown in Table 7, again the models proposed by [16], [19] and [24] achieve a high attack success rate percentage while using less than 1000 queries on the targeted model. Most of the attacks shown in this table perform well in terms of the average L_2 distance parameter.

5 DEFENCE AGAINST BLACK BOX ATTACKS

In this section we provide a detailed discussion of recent works in the defenses proposed against adversarial attacks. Common paradigms in defences follow gradient masking, input transformation, adversarial training, adversarial detection based approaches. The following sub sections discuss each of these paradigms, with key examples from each concept.

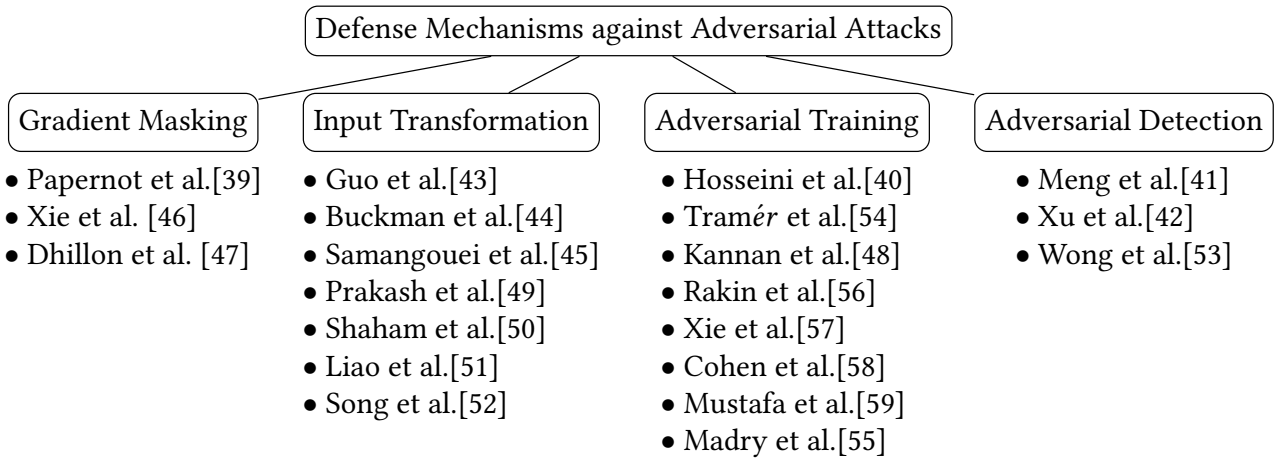
5.1 Defense Techniques

5.1.1 Gradient Masking.

Month	Year	Method	Focus/ Contribution of Research
Mar	2016	Papernot et al.[39]	Distillation
Mar	2017	Hosseini et al.[40]	NULL Labeling
Sep	2017	Meng et al.[41]	MagNet
Dec	2017	Xu et al. [42]	Feature Squeezing
Jan	2018	Guo et al.[43]	Input Transformations
Feb	2018	Buckman et al.[44]	Thermometer Encoding
Feb	2018	Samangouei et al.[45]	Defense-GAN
Feb	2018	Xie et al.[46]	Randomization
Mar	2018	Dhillon et al.[47]	SAP
Mar	2018	Kannan et al.[48]	ALP
Mar	2018	Prakash et al.[49]	Pixel Deflection
Apr	2018	Shaham et al.[50]	Basis Functions Transformation
May	2018	Liao et al.[51]	Guided Denoiser
May	2018	Song et al.[52]	Pixel Defend
Jun	2018	Wong et al.[53]	Convex Adversarial Polytope
Sep	2018	Tramér et al.[54]	Ensemble Adversarial Training
Sep	2018	Madry et al.[55]	M-PGD
Nov	2018	Rakin et al.[56]	PNI
Mar	2019	Xie et al.[57]	Feature Denoising
Jun	2019	Cohen et al.[58]	Randomized Smoothing
Jul	2019	Mustafa et al.[59]	Hidden Space Restriction

Table 8. A comparison of works that have focused on defenses against an adversarial setting.

Classification of prior work on defense strategies proposed against adversarial attacks.



- **Distillation**: The defense mechanism proposed by the authors [39] is based on *distillation* introduced by Hinton et al.[60]. Distillation in an easy sense is used to transfer knowledge from larger to smaller DNN architectures in order to reduce computational complexity. Given a training data set $\{(X, Y(X)) : X \in \mathcal{X}\}$ where X are the images and Y are the one-hot vector(e.g {0,0,1...0}) spread over the length of classes with the correct class having 1, a deep neural network F is trained with softmax output layer with temperature T which is used to

control the knowledge transfer. $F(X)$ is the probability vector over all the classes. A second training set $\{(X, F(X)) : X \in \mathcal{X}\}$ is obtained and $F(X)$ obtained from the first DNN are here used as the *soft labels*. A second model, F^d is trained using this dataset and this model is called as *distilled* model. This model is much more smoothed and the amplitude of its network gradients exploited by adversarial samples are also thereby reduced.

- **Randomization**: The goal of this paper [46] is to alleviate adversarial attacks by introducing two random layers at inference time: 1) *Random Resizing* 2) *Random Padding*. These randomization effects help defending against single-step and iterative gradient based attacks. The random resizing layer converts a image of size $W*H*3$ into a new image of size $W'*H'*3$ such that $|W'-W|$ and $|H'-H|$ are within a reasonably small range. The random padding layer pads zeros around resized image in a random manner. Suppose after random padding the size of the image is $W''*H''*3$ then there are $(W''-W'+1)*(H''-H'+1)$ different possible patterns.

- **SAP**: The authors [47] have proposed *Stochastic Activation Pruning* as a method for guarding pretrained networks against adversarial examples. During the forward pass, they stochastically prune a subset of activations in each layer. After pruning, they scale up the surviving activations to normalize the dynamic range of inputs to the subsequent layers. Deriving inspiration from game theory[61], they treat the problem as a minimax zerosum game between adversary and the model:

$$\pi^*, \rho^* := arg\ min_\pi max_\rho \mathbb{E}_{p\sim\pi, r\sim\rho}[J(M_p(\theta), x + r, y)], \tag{31}$$

where ρ is the adversary's policy from which $r \sim \rho$ perturbation is sampled and π is the defender's policy from which $p \sim \pi$ model parameters corresponding to $M_p(\theta)$ is sampled. The adversary's goal is to maximize the defender's loss under strategy ρ and defender's goal is to minimize the loss by changing model parameters θ to $M_p(\theta)$ under strategy π. Now given the i'th layer activation map,$h^i \in \mathbb{R}^{a^i}$, the probability of sampling j'th activation with value $(h^i)_j$ is given by:

$$p_j^i = \frac{|(h^i)_j|}{\sum_{k=1}^{a^i} |(h^i)_k|}$$

If an activation is sampled it is scaled up by inverse of the probability of sampling over all draws. If not, then the activation is set to 0. In this way, inverse propensity scoring of each activation is preserved. Under an instance p of policy π, r_p^i sample are drawn giving a new activation map $M_p(h^i)$:

$$M_p(h^i) = h^i \odot m_p^i \qquad (m_p^i)_j = \frac{\mathbb{I}((h^i)_j)}{1-(1-p_j^i)^{r_p^i}}$$

This scheme is similar to dropout[62] however it differs from it in its sampling technique: SAP is more likely to sample activations that are high in absolute value whereas dropout samples each activation with the same probability.

5.1.2 *Input Transformation.*

- **Input Transformations**: The authors [43] tested several input transformations-1) Image Cropping and Rescaling 2) Bit Depth Reduction[42] 3) JPEG Image Compression and Decompression[63] 4) Total Variance Minimization[64] 5) Image Quilting[65] and their corresponding effects in withstanding different adversarial attacks. Specifically, *total variance*

minimization and *image quilting* were effective defenses when the model is trained on transformed images. Their strength lies in their *non-differential nature* and *randomness* which makes it difficult for the adversary to circumvent them. Total Variance Minimization involves synthesizing a image **z** from adversarial input x by solving:

$$min_z||(1-X)\odot(z-x)||_2 + \lambda_{TV}.TV_p(\mathbf{z}) \tag{32}$$

This approach selects a small set of pixels and reconstructs the image x that is consistent with the selected pixels. The pixels are selected by sampling a Bernoulli Random Variable $X(i,j,k)$ at each pixel position (i,j,k), and the pixel is set when $X(i,j,k)=1$. $\odot$ is the element wise multiplication.

$$TV_p(\mathbf{z}) = \sum_{k=1}^{K}\left[\sum_{i=2}^{N}||\mathbf{z}(i,:,k)-\mathbf{z}(i-1,:,k)||_p + \sum_{j=2}^{N}||\mathbf{z}(:,j,k)-\mathbf{z}(:,j-1,k)||_p\right] \tag{33}$$

$TV_p(z)$ is the total l_p variation of z. The total variation(TV) measures the fine-scale variation in z, thereby encouraging the removal of adversarial perturbation during TV minimization. Image Quilting is a technique which involves synthesizing images by combining images obtained from a database. The algorithm places image patches for a predefined set of grid points and computes minimum graph cuts in overlapping boundary regions to remove edge artifacts.

- **Thermometer Encoding**: Buckman et al. [44] proposes a *input discretization* method to break the linear extrapolation behaviour of machine learning models. Goodfellow et al.[66] provided evidence that several network architectures, are vulnerable to adversarial examples as the loss function of these networks tend to be highly linear with respect to its inputs. Though neural networks in principle have capacity to represent highly nonlinear functions, networks trained via SGD tend to converge to linear solutions. One hypothesis explaining this states that the nonlinearities used in networks are piece-wise linear. Instead of using highly complex nonlinear functions which are difficult to train and generalize the authors have specified a simple non differentiable and non linear transformation(*discretization*). Two discretization methods have been specified: 1) **One Hot Encoding**: for a pixel $i \in \{1,2,...n\}$:

$$f_{oneshot}(x_i) = \chi(b(x_i)) \tag{34}$$

where $b(\theta)$ is the quantization function. For $b_i = \frac{i}{k}$, $b(\theta)$ is the largest index $\alpha \in \{1,2,...k\}$ such that $\theta \leq b_\alpha$. $\chi(j)$ is the one hot vector of j i.e $\chi(j) \in \mathbb{R}^k$ such that:

$$\chi(j)_l = \begin{cases} 1 & if\ l=j \\ 0 & otherwise \end{cases}$$

So for an input pixel i, $b(.)$ obtains the index α such that $i \leq b_\alpha$. $\chi(.)$ computes the one hot vector spread across the length of the input dimension and sets 1 at index α and 0 everywhere else.
2) **Thermometer Encoding**: For a pixel $i \in \{1,2,...k\}$:

$$f_{therm}(x_i) = \tau(b(x_i)) = \mathbb{C}(f_{oneshot}(x_i)) \tag{35}$$

where $\mathbb{C}$ is the cumulative sum function, $\mathbb{C}(c)_l = \sum_{j=0}^{l} c_l$. $\tau(j)$ is the thermometer vector i.e $\tau(j) \in \mathbb{R}^k$ such that:

$$\tau(j)_l = \begin{cases} 1 & if\ l >= j \\ 0 & otherwise \end{cases}$$

$\tau(j)$ computes the thermometer vector which is one hot vector with consecutive 1 after the index α obtained using the $b(.)$ function.

- **Defense-GAN**: Samangouei et al. [45] proposed a defense mechanism based on *Wasserstein-GAN*[67]. A GAN is a model in which min-max game is played between two adversaries:- *a generator(G)* and *a discriminator(D)*. The *W-GAN* is similar to *GAN* but instead of min-max loss, wasserstein loss is used for learning the distribution of unperturbed images. At inference time, it finds a close output to the given input image which does not contain adversarial changes i.e for a input image x we obtain z^* such that:

$$z^* = min_z ||G(z) - x||_2^2 \tag{36}$$

The above minimization problem is solved by doing L steps of Gradient Descent on R different random initializations of z.Then $G(z^*)$ is added as the input to the classifier. The motivation of the above mechanism is derived from the global optimality of *GAN* min-max loss when $p_g = p_{data}$ i.e the probability that the image is sampled from data distribution is equal to that of the generator and if G and D have enough capacity to represent the data, and the training algorithm is such that p_g converges to p_{data} then:

$$\mathbb{E}_{x \sim p_{data}} \left[min_z ||G_t(z) - x||_2 \right] \rightarrow 0 \tag{37}$$

where $\mathbb{E}$ is the expected reconstruction loss which after t steps of training converges to 0, thereby indicating minuscule performance loss on clean examples. This minimization of reconstruction loss helps in reducing the adversarial noise.

- **Pixel Deflection**: Two methods have been proposed by the authors for defending against adversarial attacks: 1) Pixel Deflection samples pixels randomly in an input image and replaces it with another randomly sampled pixel from its square neighborhood. Intuitively, Prakash et al.[49] shows that adversarial attacks have a tendency to add perturbations in the entire plane of the image disregarding the location of the object. The background pixels are not really that useful in the classification of the object. By *Pixel Deflection*, certain pixels are dropped preserving enough background for classification and mitigate the impact of adversarial attacks. The probability for a pixel to be dropped is inversely proportional to the probability of that pixel constituting the part of an object i.e

$$Pr_u(P_i) \propto \frac{1}{Pr_o(P_i)}$$

2) Wavelet Denoising is used to reduce noise introduced by both pixel deflection and adversarial attacks. Prior works[68–70] illustrate certain regularities in wavelet responses of natural images which can be used to denoise images. The author's method uses *Discrete Wavelet Transform*[71] which converts the image signal into orthornormal wavelets. These wavelets form basis for image's space, orientation and scale etc.

- **Basis Function Transformations**: Shaham et. al[50] augments images at inference time using basis functions. The transformations smoothen the images thereby reducing adversarial noise. Basis functions are known functions (usually polynomial) on which linear regression

is performed in order to model a non-linear function. The basis functions used in this paper include: 1) *Low Pass Filtering* 2) *Principal Component Analysis* 3) *Wavelet Approximation* 4) *Soft-thresholding* 5) *JPEG Compression*

- **Guided Denoiser**: Since adversarial samples are based on adding noise to clean images, the natural idea that came to the authors was to develop denoising models as a defense mechanism [51]. They defined two denoising networks:
1)**Pixel Guided Denoiser(PGD)**: Given a clean image x, the denoiser $D : x* \rightarrow \hat{x}$ is used to convert the adversarial image $x*$ into $\hat{x}$ such that $L = ||x - \hat{x}||$ is minimized where L is the L_1 norm. Two denoising functions were discussed-1) *Denoising Autoencoder(DAE)*[72] and *Denoising U-Net(DUNET)* which was DAE modified with U-Net[73].
2)**High-Level representation Guided Denoiser(HGD)**: HGD uses the same U-Net structure as DUNET but differs with it in terms of the loss function. Given a target neural network f, the HGD minimizes the loss $L = ||f(x)_l - f(\hat{x})_l||$ where $f(x)_l$ is the representation of the neural network at layer l. Three variants of HGD were also specified-1) When $l = -2$, then the representations of the top-most convolutional layer is used and the denoiser is called *Feature Guided Denoiser*. 2)When $l = -1$, logit layer representations are used and the denoiser is called *Logits Guided Denoiser*. 3)When the classification loss of target model is used as denoising loss function, then the denoiser is called *Class Guided Denoiser*.

- **PixelDefend**: The authors [52] have proposed the defense mechanism *PixelDefend* which aims to detect adversarial images and modify the input images such that it resembles more the images derived from the training distribution $p(x)$. The training distribution $p(x)$ is learnt through a PixelCNN[74](p(x)=$p_{CNN}(x)$). The probabilities of the input image, $X' \sim q(X)$ and training images, $\{X_1, X_2, ...X_N\} \sim p(X)$ are evaluated using PixelCNN and rank of $p_{CNN}(X')$ in $\{p_{CNN}(X_1), p_{CNN}(X_2), ...p_{CNN}(X_N)\}$ is used as a text-statistic:

$$T = T(X'; X_1....X_N) \triangleq \sum_{i=1}^{N} \mathbb{I}[p_{CNN}(X_i) \leq p_{CNN}(X')].$$

This is used to calculate the *p-value* which is used as a metric to decide whether image is adversarial or not:

$$p = \frac{1}{N+1}\left(\sum_{i=1}^{N} \mathbb{I}[T_i \leq T] + 1\right) = \frac{T+1}{N+1} = \frac{1}{N+1}\left(\sum_{i=1}^{N} \mathbb{I}[p_{CNN}(X_i) \leq p_{CNN}(X')] + 1\right) \quad (38)$$

Now the input image X is used to calculate the image X^* that maximizes $p(X)$ subject to the constraint that X^* is within the ϵ_{defend} ball(maximum perturbation) of X:

$$max_{X^*} p(X^*)$$
$$s.t \;\; ||X - X^*||_\infty \leq \epsilon_{defend}$$

Iteratively pixels ($x \leftarrow X[i, j, k]$) are sampled from the range $R \leftarrow [max(x-\epsilon_{defend}, 0), min(x+\epsilon_{defend}, 255)]$, with the goal to maximize the training data distribution $p(X^*)$.

5.1.3 *Adversarial Training.*

- **NULL Labeling**: In this paper [40], the authors propose to block transferability by training a classifier such that as input noise level increases, the classifier shows lower confidence on

original label and instead declares the input invalid. The method consists of three steps: 1) initial training 2) computing NULL probabilities for adversarial examples generated with different amount of perturbations 3)adversarial training.
The compute function f which is used for computing the null probability p_{NULL} for adversarial examples in this case is taken as the **MG Method**. During Adversarial training phase, the adversarial examples are generated using **STG Method**. The network is trained based on the following optimization problems:

$$\delta X^* = argmin_{\delta X} l(X + \delta X, Z_T; \theta), \quad s.t. ||\delta X||_0 \sim U[1, N_{max}]$$
$$\theta^* = argmin_\theta \alpha l(X, Z; \theta) + (1 - \alpha) l(X + \delta X^*, Z_A; \theta)$$

- **Ensemble Adversarial Training**: Tramer et al.[54] hypothesized that Kurakin et al.'s[75] adversarially trained model on single-step attacks was still vulnerable to multi-step attacks. He proposed *Ensemble Adversarial Training* as method for decoupling the adversarial examples from the parameters of the trained model and increase the diversity of perturbations seen during training. The algorithm augments the training dataset of the target model with adversarial examples transferred from *other static pre-trained* models. Intuitively, transfer of adversarial examples from multiple models acts as a good approximation of maximization problem in Madry et al.'s[55] paper which was not possible from transferable single step attacks.

- **ALP**: In this paper [48], Kannan et al. proposed a method called as *logit pairing* and its 2 variants- *clean* and *adversarial*. They compared their method with a modified version of defense proposed by Madry et al.[76] which they formulated as **mixed minibatch PGD**. **m-PGD** is stated as min-max problem:

$$argmin_\theta \left[\mathbb{E}_{(x,y)\in\hat{p}_{data}} \left(max_{\delta \in S} L(\theta, x + \delta, y) \right) + \mathbb{E}_{(x,y)\in\hat{p}_{data}} \left(L(\theta, x, y) \right) \right] \tag{39}$$

Adversarial and Clean Logit Pairing involve minimizing following losses respectively:

$$J(\mathbb{M}, \theta) + \lambda \frac{1}{m} \sum_{i=1}^{m} L \left(f(x^{(i)}; \theta), f(\bar{x}^{(i)}; \theta) \right) \tag{40}$$

$$J^{(clean)}(\mathbb{M}, \theta) + \lambda \frac{2}{m} \sum_{i=1}^{\frac{m}{2}} L \left(f(x^{(i)}; \theta), f(x^{(i+\frac{m}{2})}; \theta) \right) \tag{41}$$

Here $J(\mathbb{M}, \theta)$ is the classifier loss for minibatch $\mathbb{M}$. λ and m are the penalizing coefficient and training size respectively. $f(x, \theta)$ is the logit layer and **L** is the loss function for calculating the similarity of two images. Here **L** is the L^2 distance metric and in equation [40] $(x, \bar{x})$ are the training and its adversarial image respectively and in equation [41] $(x, \acute{x})$ are two random training images respectively.

- **M-PGD**: Madry et al. [55] conducts a careful examination of the optimization problem in the study of adversarial attacks and robust models. He proposes a guarantee that a adversarial robust model should satisfy in the form of a *Saddle-Point* (cite) problem:

$$min_\theta \rho(\theta), \quad \text{where} \quad \rho(\theta) = \mathbb{E}_{(x,y)\sim D} \left[max_{\delta \in S} L(\theta, x + \delta, y) \right] \tag{42}$$

Evaluation of local maxima of loss values by using *PGD*[37] attack on *MNIST* and *CIFAR-10* models suggest that PGD is the universal adversary among the first order approaches.

Robustness against PGD yields robustness against all first order adversaries i.e as long as the adversary uses the gradients of the loss function w.r.t input, it will not find significantly better local maxima than PGD. So in order to maximize the inner optimization problem, PGD is the best first-order attack whereas adversarial training is the solution for the outer optimization problem. The authors also provide a comparitive study of network capacity against adversarial robustness. For the weaker *black-box* attack based on transferability, increased network capacity results in decreased transferability and adversarial training using stronger adversaries also decreases transferability.

- **PNI**: Rakin et al. [56] has proposed *Parametric Noise Injection* which aims to make the model more robust by adding trainable gaussian noise at each layer on either activation or weights during the adversarial training phase:

$$\bar{v_i} = f_{PNI}(v_i) = v_i + \alpha_i.\eta; \quad \eta \sim \mathcal{N}(0, \sigma^2) \tag{43}$$

$$where \;\; \sigma = \sqrt{\frac{1}{N}\sum_i (v_i - \mu)} \tag{44}$$

Here v_i is the element of noise free tensor v, which can be input/weight/activation. η is the noise sampled from the gaussian distribution and α_i is the learnable noise parameter which is same along a layer in order to avoid over-parameterization and facilitate convergence. The gradient computation of loss w.r.t to noise is:

$$\frac{\partial \mathcal{L}}{\partial \alpha} = \sum_i \frac{\partial \mathcal{L}}{\partial f_{PNI}(v_i)} \frac{\partial f_{PNI}(v_i)}{\partial \alpha}; \quad \frac{\partial f_{PNI}(v_i)}{\partial \alpha} = \eta$$

Using a gradient descent optimizer with momentum, the optimization of α at step j can be written as:

$$V_i^j = m.V_i^{j-1} + \frac{\partial \mathcal{L}^{j-1}}{\partial \alpha}; \quad \alpha^j = \alpha^{j-1} - \epsilon.V_i^j$$

where m is the momentum, ϵ is the learning rate and V is the updating velocity.

- **Feature Denoising**: Xie et al. [57] suggests that small perturbations added in the input domain lead to substantial noise in the feature maps of the network. While the feature maps of clean images focus on semantically relevant content, the feature maps of adversarial images are even activated in irrelevant regions.
  Motivated by this, he proposes *feature denoising* as a defense mechanism. Adversarial Training of the classifier is performed together with the addition of a *denoising block*. The structure of the block is inspired from *self-attention*[77] and *Non-Local Networks*[78].
  Several denoising operations have been mentioned in this paper: 1) **Non-Local Means**[79]: For a input feature map x, a denoised feature map y is computed by weighted mean of features in all spatial locations $\mathcal{L}$:

$$y_i = \frac{1}{C(x)} \sum_{\forall j \in \mathcal{L}} f(x_i, x_j).x_j, \tag{45}$$

where $f(x_i, x_j)$ is the feature dependent weighting function and $C(x)$ is the normalization function. Two Variants have been discussed- 1) *Gaussian(softmax)* sets where $f(x_i, x_j) = e^{\frac{1}{\sqrt{d}}\theta(x_i)^T\phi(x_j)}$ and $C = \sum_{\forall j \in \mathcal{L}} f(x_i, x_j)$ and 2) *Dot Product* sets where $f(x_i, x_j) = x_i^T x_j$ and

$C(x) = N$ where N is the number of pixels in x. 2) **Bilateral Filter**[80]:It is similar to Equation[45] but differs from it in its neighborhood $\Omega(i)$, which is a local region (e.g a 3X3 patch) around pixel i. 3) **Mean Filter**: Simplest denoising operation is the mean filter i.e average pooling with a stride of 1. Mean filters reduce noise but also smooth structures. 4) **Median Filter**: It is defined as $y_i = median\{\forall j \in \Omega(i) : x_j\}$ where the median is over a local region $\Omega(i)$ and is performed separately for each channel. It is good at removing *salt and pepper* noise.

- **Randomized Smoothing**: The goal of this defense method is to obtain a new classifier g from base classifier f which returns the most probable class when the input x perturbed with Gaussian Noise is added to the base classifier:

$$g(x) = argmax_{c \in \mathcal{Y}} \mathbb{P}(f(x + \epsilon) = c), \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I) \tag{46}$$

Now n samples of Gaussian noise is drawn and added to the input image x. These noisy samples of x are given as input to the base classifier f and the class counts are obtained. If $\hat{c_A}$ and $\hat{c_B}$ are the two classes with maximum counts n_A and n_B then a Binomial Hypothesis test is performed to obtain the p-value ($n_A \sim$ Binomial($n_A + n_B$,p)). If the p-value is less than a constant α then return $\hat{c_A}$ otherwise abstain from returning an output.

- **Hidden Space Restriction**: The authors of this paper suggest that the main reason for perturbations is the close-proximity of different class samples in the learned space. To counter this, they proposed to class-wise disentangle intermediate feature representation of deep networks i.e force the features of each class to lie inside a convex polytope which are maximally seperated from the polytopes of other classes. For a input sample x let the adversarial polytope be $P_\epsilon(x;\theta) = F_\theta(x + \delta)\ s.t., ||\delta||_p \leq \epsilon$, then $O_\epsilon^{i,j}$ represents the overlap region between data pair (i, j) i.e $O_\epsilon^{i,j} = P_\epsilon(x_{y_i}^i;\theta) \cap P_\epsilon(x_{y_j}^j;\theta)$. The authors propose that reducing this overlap region between differnt classes will result in lower adversarial success. In order to achieve this the authors augment the loss function used while training. This new loss added is called *Prototype Conformity Loss*:

$$\mathcal{L}_{PC}(x, y) = \sum_i \left\{ ||f_i - w_{y_i}^c||_2 - \frac{1}{k-1} \sum_{j \neq y_i} \left(||f_i - w_j^c||_2 + ||w_{y_i}^c - w_j^c||_2 \right) \right\} \tag{47}$$

where f_i denotes the DNN output representation of image i and $w_{y_i}^c$ are the trainable centroids of class y_i. At inference time, the label is given to an image i as follows: $\hat{y}_i = argmin_j ||f_i - w_{y_j}^c||$. The total loss function for training the model is:

$$\mathcal{L}(x, y) = \mathcal{L}_{CE}(x, y) + \mathcal{L}_{PC}(x, y) \tag{48}$$

where $\mathcal{L}_{CE}(.)$ is the cross entropy loss. In order to achieve a similar effect in intermediate layers, $\mathcal{L}_{PC}(.)$ is calculated using a branch $\mathcal{G}_\phi$ which maps the features to a lower dimension. So the final loss becomes:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{CE} + \sum_l^L \mathcal{L}_{PC} \\ where \quad & \mathcal{L}_{PC}^l = \mathcal{L}_{PC}(f^l, y) \\ s.t., \quad & f^l = \mathcal{G}_\phi^l(F_\theta^l(x)) \end{aligned}$$

5.1.4 Adversarial Detection.

- **MagNet**: *MagNet* [41] is used for defending neural network classifiers against adversarial examples. MagNet consists of two components: 1) a *detector* which rejects examples far from the decision boundary of different classes. 2) a *reformer* which given an input x strives to find x' such that x' is a close approximation of x and also is close to the normal examples manifold(decision boundary).
Detector: The detector is the function $d : \mathbb{S} \rightarrow 0, 1$ which detects whether the input is adversarial. Two kinds of detector were mentioned in the literature: 1) *Detector based on Reconstruction Error* - In this an autoencoder is trained using the normal samples. The loss function minimized in this case is:

$$L(\mathbb{X}_{train}) = \frac{1}{|\mathbb{X}_{train}|} \sum_{x \in \mathbb{X}_{train}} ||x - ae(x)||_2$$

$ae(x)$ denotes the reconstructed image obtained from the autoencoder ae. The reconstruction error on a test sample x is:

$$E(x) = ||x - ae(x)||_p \tag{49}$$

where $E(x)$ is the l_p distance between input image and its reconstructed image. If $E(x)$ is greater than some constant δ_p then the input is an adversarial example otherwise it is not. 2) *Detector based on probability divergence* - Let $f(x)$ be the output of the last layer(softmax) of the neural network f given an input x. Then in this case probability divergence of $f(x)$ and $f(ae(x))$ is evaluated using Jensen-Shannon Divergence:

$$JSD(P||Q) = \frac{1}{2} D_{KL}(P||M) + \frac{1}{2} D_{KL}(Q||M) \tag{50}$$

where $D_{KL}(P||Q) = \sum_i P(i) log \frac{P(i)}{Q(i)}$ is the Kullback–Leibler Divergence and $M = \frac{1}{2}(P+Q)$. The Jensen–Shannon divergence is a method of measuring the similarity between two probability distributions P and Q.

Reformer: The reformer is the function $r : \mathbb{S} \rightarrow \mathbb{N}_t$ which tries to reconstruct the test input so that it is closer to the training data manifold. An ideal reformer: 1) should not change classification results of normal samples and 2) should change adversarial samples adequately such that the reformed samples are close to normal samples. Two kinds of reform strategies were proposed: 1) *Noise based Reformer*: a naive reformer which adds random noise to the input sampled from a Gaussian distribution i.e $r(x) = clip(x + \epsilon.y)$ where $y \sim N(y; 0, I)$. 2) *Autoencoder based Reformer*: In this case the autoencoder trained for detector $ae(x)$ is used. For an input x' the reconstruction obtained from autoencoder $ae(x')$ is fed to the target classifier as it is closer to the normal samples manifold.

- **Feature Squeezing**: Xu et al. [42] argues that feature input spaces are unnecessarily large. So he proposed a mechanism to reduce the degrees of freedom available to an adversary by squeezing out unnecessary input features. *Feature Squeezing* is a simple adversarial example detection framework that compares model's prediction on the original sample with its prediction after squeezing. If the difference is larger than a threshold value then the sample is adversarial and is discarded.
Two techniques have been proposed by the authors:1)*Squeezing Color Bits*: Grayscale images comprise of 8 bit pixel values whereas color images consist of 24 bit pixels. The bit depth

can be considerably reduced without much loss of information thereby reducing the search space for adversarial samples. This is done by using binary filter with 0.5 as the cutoff.
2) *Spatial Smoothing*: It is a image processing technique that is used to reduce noise. Its two variations have been mentioned:1) *Local Smoothing*: Smoothes the value of pixel in a particular window size by using nearby pixels.In this paper, Median Smoothing was used.
2) *Non Local Smoothing*:For a given image patch, non local smoothing finds several similar patches in a larger area and then replaces the center patch with the average of those patches.

- **Convex adversarial polytope**: In this paper, Wong et al. [53] proposed a method for obtaining *provably robust* ReLU classifiers. For a ReLU based neural network f where $f_\theta : \mathbb{R}^{|x|} \rightarrow \mathbb{R}^{|y|}$, we obtain an adversarial polytope ($\mathcal{Z}_\epsilon(x)$) which is the set of all final layer activations that can be obtained by perturbing the input with Δ. Here, Δ is l_∞ norm bounded by ϵ:

$$\mathcal{Z}_\epsilon(x) = \{f_\theta(x + \Delta) : ||\Delta||_\infty \leq \epsilon\} \tag{51}$$

Now, a *convex outer bound* on this adversarial polytope is created. If no point inside this *outer bound* changes the class label of the input then no label is changed in the true adversarial polytope as well. This bound is obtained by solving the following linear program:

$$z \geq 0,\ z \geq \hat{z},\ -u\hat{z} + (u - l)z \leq -ul \tag{52}$$

Here, $u, l, \hat{z}, z$ are the lower bound, upper bound, pre ReLU activation and ReLU activation respectively. Given a dataset $(x_i, y_i)_{i=1,2,\ldots,N}$, instead of training the model using these data points, the model is trained using the farthest points (i.e with the highest loss) from x_i which are l_∞ by ϵ:

$$minimize_\theta \sum_{i=1}^{N} max_{||\Delta||_\infty \leq \epsilon} L(f_\theta(x_i + \Delta), y_i) \tag{53}$$

L is the loss function and θ are the parameters of the feedforward ReLU neural network.

5.2 Comparitive Analysis of Defense Techniques

Similar to the attack section, the defense techniques have now been tested and compared on three datasets i.e. MNIST, CIFAR-10 and ImageNet. The results have been compared along following dimensions: the attack model on which the defense technique was used, classification accuracy (CA) which is the accuracy obtained on clean images, attack success rate (ASR) which highlights the performance of the specific attack model and the specific parameters used. CA and ASR have two subcategories as well: with defense and without defense which measure the above mentioned metrics with and without defense techniques. The results have been obtained from the respective research studies.

Method	Attack Model	Classification Accuracy		Attack Success Rate		Specific Parameters Used
		Without Defense	With Defense	Without Defense	With Defense	
Papernot[39]	Limitation of Deep learning model	99.51%	99.05%	95.89%	0.45%	Temperature T = 100
Hosseini[40]	FGSM	99.35%	99.39%	90%	45%	α = 0.5, label Smoothing Parameter q = 0.9
Meng[41]	FGSM	99.40%	99.10%	8.90%	0%	L_∞ metric, ϵ = 0.010, epochs = 100, lr = 0.001
	Iterative	99.40%	99.10%	28%	0%	L_∞ metric, ϵ = 0.010, epochs = 100, lr = 0.001
	Deepfool	99.40%	99.10%	80.90%	0.60%	L_∞ metric, epochs = 100, lr = 0.001
	C&W	99.40%	99.10%	100%	0.20%	L_∞ metric, epochs = 100, lr = 0.001
Xu[42]	FGSM	99.43%	99.28%	46%	8%	L_∞, Bit Depth Method of 1 bit
	C&W	99.43%	99.28%	100%	0%	L_∞, Bit Depth Method of 1 bit
	JSMA	99.43%	99.28%	71%	18%	L_0 metric, Median Smoothing of 3x3
Buckman[44]	FGSM	99.30%	98.67%	100%	3.71%	l_∞ metric, ϵ = 0.3, 40 steps for iterative attacks
	PGD/LS-PGA	99.30%	98.67%	100%	5.70%	l_∞ metric, ϵ = 0.3, 40 steps for iterative attacks, ξ = 0.01, δ = 1.2
Samangouie[45]	FGSM	99.70%	-	71.84%	8.95%	ϵ = 0.3, L = 200, R = 10
Kannan[48]	PGD	-	98.80%	-	3.60%	l_∞ metric, ϵ = 0.3, lr = 0.045
	Madry Black Box	-	98.80%	-	2.50%	l_∞ metric, ϵ = 0.3, lr = 0.045
Wong[53]	FGSM	98.20%	94.18%	50.01%	3.93%	l_∞ metric, ϵ = 0.1
	PGD	98.20%	94.18%	81.68%	4.11%	l_∞ metric, ϵ = 0.1
Tramer[54]	FGSM	-	99.30%	72.80%	14%	l_∞ metric, ϵ = 0.3, lr = 0.045
Madry[76]	FGSM	99.20%	98.80%	58.10%	4.40%	l_∞ metric, ϵ = 0.3
	PGD	99.20%	98.80%	74%	10.70%	l_∞ metric, ϵ = 0.3, 40 iterations of PGD
Mustafa[59]	FGSM	98.71%	99.53%	77%	21.70%	epochs = 50, lr = 0.1, ϵ = 0.3
	C&W	98.71%	99.53%	79.10%	22.80%	c = 10, 1000 iteration steps for C&W
	PGD	98.71%	99.53%	88.10%	30.50%	10 iteration steps for PGD with step size $\frac{\epsilon}{10}$

Table 9. Defense techniques against attacks on MNIST data set.

Method	Attack Model	Classification Accuracy		Attack Success Rate		Specific Parameters Used
		Without Defense	With Defense	Without Defense	With Defense	
Papernot[39]	-	80.95%	81.34%	87.59%	5.11%	Temperature T = 100
Meng[41]	FGSM	90.60%	86.80%	54%	0.10%	L_∞ metric, ϵ = 0.050, epochs = 400, lr = 0.001
	Iterative	90.60%	86.80%	88.90%	0.10%	L_∞ metric, ϵ = 0.025, epochs = 400, lr = 0.001
	Deepfool	90.60%	86.80%	95.50%	6.60%	L_∞ metric, epochs = 400, lr = 0.001
	C&W	90.60%	86.80%	100%	17%	L_∞ metric, epochs = 400, lr = 0.001
Xu[42]	FGSM	94.84%	89.29%	85%	62%	L_∞ metric, Median Smoothing of 2x2
	C&W	94.84%	89.29%	100%	16%	L_∞ metric, Median Smoothing of 2x2
	Deepfool	94.84%	89.29%	98%	17%	L_2 metric, Median Smoothing of 2x2
	JSMA	94.84%	89.29%	71%	16%	L_0 metric, Median Smoothing of 2x2
Buckman[44]	FGSM	94.29%	87.67%	56.11%	19.04%	l_∞ metric, ϵ = 0.031, 7 steps for iterative attacks
	PGD/LS-PGA	94.29%	87.67%	98.34%	20.84%	l_∞ metric, ϵ = 0.031, 7 steps for iterative attacks, ξ = 0.01, δ = 1.2
Dhillon[47]	FGSM	89.80%	-	30%	22%	epochs = 150, Initial lr = 0.5
	I-FGSM	89.80%	-	40%	15%	λ = 1, SAP-100
Song[52]	FGSM	92%	88%	89%	76%	l_∞ metric, ϵ_{attack} = 2/8/16, ϵ_{defend} = 16
	Deepfool	92%	88%	94%	20%	l_∞ metric, ϵ_{attack} = 2/8/16, ϵ_{defend} = 16
	C&W	92%	88%	100%	22%	l_∞ metric, ϵ_{attack} = 2/8/16, ϵ_{defend} = 16
Rakin[56]	FGSM	92.11%	84.78%	85.92%	45.96%	ϵ = 0.3/1 and ϵ = 8/255 under l_2 and l_∞ distance respectively
	PGD	92.11%	84.78%	100%	54.17%	Number of iteration steps is 7 for PGD
Cohen[58]	Deepfool	90%	78%	80%	35%	l_2 metric, α = 0.001, n = 100000
Madry[76]	FGSM	95.20%	90.30%	67.30%	4.90%	l_∞ metric, ϵ = 8
	PGD	95.20%	90.30%	96.50%	54.20%	l_∞ metric, ϵ = 8, 7 steps of size 2 for PGD
Mustafa[59]	FGSM	90.80%	90.45%	61%	14.50%	epochs = 200, lr = 0.1, ϵ = 0.03
	C&W	90.80%	90.45%	68.20%	16.70%	c = 0.1, 1000 iteration steps for C&W
	PGD	90.80%	90.45%	70.90%	23.60%	10 iteration steps for PGD with step size $\frac{\epsilon}{10}$

Table 10. Defense techniques against attacks on CIFAR-10 data set.

Method	Attack Model	Classification Accuracy		Attack Success Rate		Specific Parameters Used
		Without Defense	With Defense	Without Defense	With Defense	
Xu[42]	FGSM	69.70%	62.10%	99%	67%	L_∞ metric, Median Smoothing of 3x3
	C&W	69.70%	62.10%	99%	23%	L_∞ metric, Non Local Means Method (11-3-4)
	DeepFool	69.70%	62.10%	89%	28%	L_2 metric, Median Smoothing of 2x2
Guo[43]	FGSM	-	75.10%	-	29.63%	L_2 metric
	I-FGSM	-	75.10%	-	28.48%	Pixel dropout probability p = 0.5
	DeepFool	-	75.10%	-	28.53%	Total Variance minimizer $\lambda_{TV} = 0.03$
	C&W	-	75.10%	-	29.50%	Quilting Patch Size of 5x5
Xie[46]	FGSM	100%	97.30%	67%	47.60%	l_∞ metric, $\epsilon = 10$
	DeepFool	100%	97.30%	100%	1.70%	l_∞ metric, $\epsilon = 10$
	C&W	100%	97.30%	100%	3.10%	l_∞ metric, $\epsilon = 10$
Kannan[48]	PGD	-	-	-	72.10%	l_∞ metric, $\epsilon = 0.3$, lr = 0.045
	Madry Black Box	-	-	-	53.30%	l_∞ metric, $\epsilon = 0.3$, lr = 0.045
Prakash[49]	FGSM	100%	98.90%	80%	18.50%	l_2 metric, $\epsilon = 0.02 - 0.04$
	I-FGSM	100%	98.90%	85.90%	16.30%	Coefficient for BayesShrink, $\sigma = 0.04$
	DeepFool	100%	98.90%	73.70%	9.70%	Window size for pixel deflection, r = 10
	JSMA	100%	98.90%	74.50%	3%	Number of pixel deflections, K = 100
	C&W	100%	98.90%	95.20%	2%	ResNet-50 Model
Liao[51]	FGSM-White Box	76.70%	75.30%	85.60%	63.30%	epochs = 20 to 30, initial lr = 0.001, l_∞ metric, $\epsilon = 4/16$
	FGSM-Black Box	76.70%	75.30%	59%	43.30%	epochs = 20 to 30, initial lr = 0.001, l_∞ metric, $\epsilon = 4/16$
Tramer[54]	best(FGSM, I-FGSM, PGD)	80.40%	79.80%	44.40%	27%	l_∞ metric, $\epsilon = 16$, lr = 0.045
Xie[57]	NIPS '17 CAAD	78.91%	79.08%	100%	50.50%	l_∞ metric, $\epsilon = 32$, $\alpha = 1$
Cohen[58]	DeepFool	78%	70%	95%	49%	l_2 metric, $\alpha = 0.001$, n = 100000

Table 11. Defense techniques against attacks on ImageNet data set.

6 CONCLUSION

Our primary motive behind presenting this study is to encourage the community to push forward research in the black-box discipline. Analysis of the defense techniques discussed in this study reveals that majority of these methods have been tested mostly on white-box adversarial attacks. To counter the security challenges put forward by the black-box adversaries, defense mechanisms should also be tested comprehensively on these models to enhance their robustness and applicability in a large number of scenarios.

REFERENCES

[1] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *CoRR*, abs/1312.6199, 2013.

[2] Marco Barreno, Blaine Nelson, Anthony D Joseph, and J Doug Tygar. The security of machine learning. *Machine Learning*, 81(2):121–148, 2010.

[3] Justin Gilmer, Ryan P Adams, Ian Goodfellow, David Andersen, and George E Dahl. Motivating the rules of the game for adversarial example research. *arXiv preprint arXiv:1807.06732*, 2018.

[4] Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430, 2018.

[5] Qiang Liu, Pan Li, Wentao Zhao, Wei Cai, Shui Yu, and Victor CM Leung. A survey on security threats and defensive techniques of machine learning: A data driven view. *IEEE access*, 6:12103–12117, 2018.

[6] Atul Kumar and Sameep Mehta. A survey on resilient machine learning. *arXiv preprint arXiv:1707.03184*, 2017.

[7] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, ASIA CCS '17, pages 506–519, New York, NY, USA, 2017. ACM.

[8] Nina Narodytska and Shiva Prasad Kasiviswanathan. Simple black-box adversarial perturbations for deep networks. *CoRR*, abs/1612.06299, 2016.

[9] Nina Narodytska and Shiva Kasiviswanathan. Simple black-box adversarial attacks on deep neural networks. *CoRR*, pages 1310–1318, 07 2017.

[10] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, AISec '17, pages 15–26, New York, NY, USA, 2017. ACM.

[11] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv preprint arXiv:1712.04248*, 2017.

[12] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. *CoRR*, abs/1804.08598, 2018.

[13] Minhao Cheng, Thong Le, Pin-Yu Chen, Jinfeng Yi, Huan Zhang, and Cho-Jui Hsieh. Query-efficient hard-label black-box attack: An optimization-based approach. *CoRR*, abs/1807.04457, 2018.

[14] Arjun Nitin Bhagoji, Warren He, Bo Li, and Dawn Song. Practical black-box attacks on deep neural networks using efficient query mechanisms. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[15] Yali Du, Meng Fang, Jinfeng Yi, Jun Cheng, and Dacheng Tao. Towards query efficient black-box attacks: An input-free perspective. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*, AISec '18, pages 13–24, New York, NY, USA, 2018. ACM.

[16] Chun-Chen Tu, Pai-Shun Ting, Pin-Yu Chen, Sijia Liu, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh, and Shin-Ming Cheng. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. *CoRR*, abs/1805.11770, 2018.

[17] Yucheng Shi, Siyu Wang, and Yahong Han. Curls & whey: Boosting black-box adversarial attacks. *CoRR*, abs/1904.01160, 2019.

[18] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. *CoRR*, abs/1904.02884, 2019.

[19] Jinyin Chen, Mengmeng Su, Shijing Shen, Hui Xiong, and Haibin Zheng. POBA-GA: perturbation optimized black-box adversarial attacks via genetic algorithm. *CoRR*, abs/1906.03181, 2019.

[20] Thomas Brunner, Frederik Diehl, Michael Truong Le, and Alois Knoll. Guessing smart: Biased sampling for efficient black-box adversarial attacks. *arXiv preprint arXiv:1812.09803*, 2018.

[21] Yandong Li, Lijun Li, Liqiang Wang, Tong Zhang, and Boqing Gong. NATTACK: Learning the distributions of adversarial examples for an improved black-box attack on deep neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3866–3876, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[22] Seungyong Moon, Gaon An, and Hyun Oh Song. Parsimonious black-box adversarial attacks via efficient combinatorial optimization. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4636–4645, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[23] Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Prior convictions: Black-box adversarial attacks with bandits and priors. *CoRR*, 07 2018.

[24] Moustafa Alzantot, Yash Sharma, Supriyo Chakraborty, Huan Zhang, Cho-Jui Hsieh, and Mani B. Srivastava. Genattack: practical black-box attacks with gradient-free optimization. In *Proceedings of the Genetic and Evolutionary Computation Conference, GECCO 2019, Prague, Czech Republic, July 13-17, 2019*, pages 1111–1119, 2019.

[25] Chuan Guo, Jacob R. Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Q. Weinberger. Simple black-box adversarial attacks. *CoRR*, abs/1905.07121, 2019.

[26] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. *CoRR*, abs/1608.04644, 2016.

[27] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[28] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 39–57, 2017.

[29] P.D. Lax and M.S. Terrell. *Calculus With Applications*. Undergraduate Texts in Mathematics. Springer New York, 2013.

[30] Yurii Nesterov and Vladimir G. Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

[31] Sijia Liu, Jie Chen, Pin-Yu Chen, and Alfred Hero. Zeroth-order online alternating direction method of multipliers: Convergence analysis and applications. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 288–297, Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018. PMLR.

[32] Yining Wang, Simon Du, Sivaraman Balakrishnan, and Aarti Singh. Stochastic zeroth-order optimization in high dimensions. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1356–1365, Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018. PMLR.

[33] Ken Perlin. An image synthesizer. In *Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '85, pages 287–296, New York, NY, USA, 1985. ACM.

[34] Daan Wierstra, Tom Schaul, Tobias Glasmachers, Yi Sun, Jan Peters, and Jürgen Schmidhuber. Natural evolution strategies. *J. Mach. Learn. Res.*, 15(1):949–980, January 2014.

[35] Chuan Guo, Jared S. Frank, and Kilian Q. Weinberger. Low frequency adversarial perturbation. *CoRR*, abs/1809.08758, 2018.

[36] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[37] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *CoRR*, abs/1706.06083, 2017.

[38] Uriel Feige, Vahab S. Mirrokni, and Jan Vondrák. Maximizing non-monotone submodular functions. *SIAM J. Comput.*, 40(4):1133–1153, July 2011.

[39] Nicolas Papernot, Patrick D. McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *IEEE Symposium on Security and Privacy, SP 2016, San Jose, CA, USA, May 22-26, 2016*, pages 582–597, 2016.

[40] Hossein Hosseini, Yize Chen, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. Blocking transferability of adversarial examples in black-box learning systems. *CoRR*, abs/1703.04318, 2017.

[41] Dongyu Meng and Hao Chen. Magnet: A two-pronged defense against adversarial examples. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17, pages 135–147, New York, NY, USA, 2017. ACM.

[42] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *CoRR*, abs/1704.01155, 2017.

[43] Chuan Guo, Mayank Rana, Moustapha Cissé, and Laurens van der Maaten. Countering adversarial images using input transformations. *CoRR*, abs/1711.00117, 2017.

[44] Jacob Buckman, Aurko Roy, Colin Raffel, and Ian Goodfellow. Thermometer encoding: One hot way to resist adversarial examples. In *International Conference on Learning Representations*, 2018.

[45] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. *CoRR*, abs/1805.06605, 2018.

[46] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. In *International Conference on Learning Representations*, 2018.

[47] Guneet S. Dhillon, Kamyar Azizzadenesheli, Jeremy D. Bernstein, Jean Kossaifi, Aran Khanna, Zachary C. Lipton, and Animashree Anandkumar. Stochastic activation pruning for robust adversarial defense. In *International Conference on Learning Representations*, 2018.

[48] Harini Kannan, Alexey Kurakin, and Ian J. Goodfellow. Adversarial logit pairing. *CoRR*, abs/1803.06373, 2018.

[49] Aaditya Prakash, Nick Moran, Solomon Garber, Antonella DiLillo, and James Storer. Deflecting adversarial attacks with pixel deflection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[50] Uri Shaham, James Garritano, Yutaro Yamada, Ethan Weinberger, Alex Cloninger, Xiuyuan Cheng, Kelly Stanton, and Yuval Kluger. Defending against adversarial images using basis functions transformations, 2018.

[51] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[52] Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. In *International Conference on Learning Representations*, 2018.

[53] J. Zico Kolter and Eric Wong. Provable defenses against adversarial examples via the convex outer adversarial polytope. *CoRR*, abs/1711.00851, 2017.

[54] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*, 2018.

[55] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *CoRR*, abs/1706.06083, 2017.

[56] Zhezhi He, Adnan Siraj Rakin, and Deliang Fan. Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[57] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L. Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[58] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1310–1320, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[59] Aamir Mustafa, Salman Khan, Munawar Hayat, Roland Goecke, Jianbing Shen, and Ling Shao. Adversarial defense by restricting the hidden space of deep neural networks. *International Conference on Computer Vision (ICCV)*, 2019.

[60] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015.
[61] Martin J. Osborne and Ariel Rubinstein. *A Course in Game Theory*, volume 1 of *MIT Press Books*. The MIT Press, March 1994.
[62] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.
[63] Gintare Karolina Dziugaite, Zoubin Ghahramani, and Daniel M. Roy. A study of the effect of JPG compression on adversarial images. *CoRR*, abs/1608.00853, 2016.
[64] Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259 – 268, 1992.
[65] Alexei A. Efros and William T. Freeman. Image quilting for texture synthesis and transfer. *Proceedings of SIGGRAPH 2001*, pages 341–346, August 2001.
[66] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.
[67] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan, 2017.
[68] S. G. Chang, Bin Yu, and M. Vetterli. Adaptive wavelet thresholding for image denoising and compression. *IEEE Transactions on Image Processing*, 9(9):1532–1546, Sep. 2000.
[69] David J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am. A*, 4(12):2379–2394, Dec 1987.
[70] Eero Simoncelli, P Muller, and B Vidakovic. Bayesian denoising of visual images in the wavelet domain. *CoRR*, 02 2000.
[71] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies. Image coding using wavelet transform. *IEEE Transactions on Image Processing*, 1(2):205–220, April 1992.
[72] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 1096–1103, New York, NY, USA, 2008. ACM.
[73] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.
[74] Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, koray kavukcuoglu, Oriol Vinyals, and Alex Graves. Conditional image generation with pixelcnn decoders. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4790–4798. Curran Associates, Inc., 2016.
[75] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *CoRR*, 2017.
[76] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.
[77] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017.
[78] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. *CVPR*, 2018.
[79] A. Buades, B. Coll, and J. . Morel. A non-local algorithm for image denoising. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 60–65 vol. 2, June 2005.
[80] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *Proceedings of the Sixth International Conference on Computer Vision*, ICCV '98, pages 839–, Washington, DC, USA, 1998. IEEE Computer Society.

Afrika Matematika
https://doi.org/10.1007/s13370-019-00746-4

α-Bernstein–Kantorovich operators

Naokant Deo[1] · **Ram Pratap**[1]

Received: 13 July 2018 / Accepted: 4 November 2019
© African Mathematical Union and Springer-Verlag GmbH Deutschland, ein Teil von Springer Nature 2019

Abstract
In this paper we proposed the Kantorovich form of α-Bernstein operators introduced by Chen et al. (Math Anal Appl 450:244–261, 2017). We give some auxiliary properties and study the direct local approximation theorem, Voronovskaya type asymptotic and function of bounded variation for α-Bernstein–Kantorovich operators.

Keywords Bernstein operators · Kantorovich operators · Modulus of continuity · Direct results

Mathematics Subject Classification 41A25 · 41A36

1 Introduction

Recently, Chen et al. [2] defined α-Bernstein operators as a function $h \in C[0, 1]$, any fixed real α and for $n \in \mathbb{N} = \{1, 2, \ldots\}$

$$B_{n,\alpha}(h; x) = \sum_{s=0}^{n} h_s \, p_{n,s}^{(\alpha)}(x), \quad x \in [0, 1], \tag{1.1}$$

where $h_k = h\left(\frac{k}{n}\right)$. For $n > 2$ the α-Bernstein polynomial $p_{n,s}^{(\alpha)}(x)$ of degree n is defined by $p_{1,0}^{(\alpha)}(x) = 1 - x$, $p_{1,1}^{(\alpha)}(x) = x$, and

$$p_{n,s}^{(\alpha)}(x) = \left[\binom{n-2}{s}(1-\alpha)\,x + \binom{n-2}{s-2}(1-\alpha)\,(1-x) + \binom{n}{k}\alpha x\,(1-x)\right] x^{s-1}(1-x)^{n-s-1}, \quad x \in [0, 1],$$

✉ Naokant Deo
naokantdeo@dce.ac.in

Ram Pratap
rampratapiitr@gmail.com

[1] Department of Applied Mathematics, Formerly Delhi College of Engineering, Delhi Technological University, Bawana Road, Delhi 110042, India

and studied another interesting proof of the Weierstrass Approximation Theorem. They also studied its fundamental properties, the rate of convergence and Voronovskaya type asymptotic estimate formula. When $\alpha = 1$, the α-Bernstein polynomial reduces to the classical Bernstein polynomial.

Very recently, Dhamija et al. [4,5] studied Kantorovich form of generalized positive linear operators and of modified Szász–Mirakyan operators and established direct results, convergence with the help of local and weighted approximation of proposed operators. Kantorovich variant of Bernstein operators have been given and studied by several mathematician cases, see [3,6–10].

Now we consider the Kantorovich form of α-Bernstein operators [9] as:

$$\hat{K}_{n,\alpha}(h;x) = (n+1)\sum_{s=0}^{n} p_{n,s}^{(\alpha)}(x)\int_{\frac{s}{n+1}}^{\frac{s+1}{n+1}} h(t)\,dt. \tag{1.2}$$

Very recently, Acar et. al. [1] introduced the genuine Bernstein–Durrmeyer type operators with parameter α. They studied several approximation results and also shown graphically. The aim of this paper to give some basic properties and study the direct local approximation theorem, Voronovskaya type asymptotic estimate formula and bounded variation for α-Bernstein–Kantorovich operators (1.2).

2 Preliminary results

In this section, we gave some auxiliary results which help us to prove main results.

Lemma 2.1 [9] *For $0 \le x \le 1$ and $n \in \mathbb{N}$, we have*

(i) $\hat{K}_{n,\alpha}(1;x) = 1$;
(ii) $\hat{K}_{n,\alpha}(t;x) = \frac{2nx+1}{2(n+1)}$;
(iii) $\hat{K}_{n,\alpha}\left(t^2;x\right) = \frac{1}{3(n+1)^2}\left[3\left\{\left(n^2-n-2(1-\alpha)\right)x^2 + (4n+2(1-\alpha))x\right\}+1\right]$.

Lemma 2.2 [9] *The mth moments are as follows:*

$$\mu_n^m(x) = \hat{K}_{n,\alpha}\left((t-x)^m;x\right),$$

then we have

(i) $\mu_n^1(x) = \frac{1-2x}{2(n+1)}$;
(ii) $\mu_n^2(x) = \frac{1}{3(n+1)^2}\left\{3(-n+2\alpha-1)x^2 + 3(n-2\alpha+1)x + 1\right\}$;
(iii) $\mu_n^4(x) = \frac{1}{5(n+1)^4}\left\{5\left(3n^2-4(2+3\alpha)n-(131-132\alpha)\right)x^4\right.$
$+10\left(-3n^2+4(2+3\alpha)n+(131-132\alpha)\right)x^3$
$\left.+5\left(3n^2-(13+12\alpha)n-2(80-81\alpha)\right)x^2+5(5n+(33-32\alpha))x+1\right\}$.

Lemma 2.3 *If n is sufficiently large then, we have*

(i) $\lim_{n\to\infty} n\mu_n^1(x) = \frac{1-2x}{2}$;
(ii) $\lim_{n\to\infty} n\mu_n^2(x) = x(1-x)$;
(iii) $\lim_{n\to\infty} n^2\mu_n^4(x) = 3(x(1-x))^2$.

Lemma 2.4 *From Lemma 2.2, we get the following inequalities:*

(i) $\hat{K}_{n,\alpha}\left((t-x);x\right) \leq \frac{1}{2(n+1)}$;
(ii) $\hat{K}_{n,\alpha}\left((t-x)^2;x\right) \leq \frac{1}{4(n+1)}$.

Proof We have

$$\hat{K}_{n,\alpha}\left((t-x);x\right) = \frac{(1-2x)}{2(n+1)}.$$

The maximum value of $(1-2x)$ is 1 on $[0,1]$, therefore

$$\hat{K}_{n,\alpha}\left((t-x);x\right) \leq \frac{1}{2(n+1)}.$$

Also

$$\hat{K}_{n,\alpha}\left((t-x)^2;x\right) \leq \frac{1}{(n+1)}x(1-x) \leq \frac{1}{4(n+1)}.$$

Value of $x(1-x)$ is $\frac{1}{4}$. □

For our convenience the alternative form of the operators (1.2) may be written as:

$$\hat{K}_{n,\alpha}\left(h(t);x\right) = \int_0^1 U_{n,\alpha}\left(x;t\right)h(t)dt, \tag{2.1}$$

where

$$U_{n,\alpha}\left(x,t\right) = (n+1)\sum_{s=0}^{n} p_{n,s}^{(\alpha)}\left(x\right)\chi_{n,s}\left(t\right),$$

and $\chi_{n,s}\left(t\right)$ is the characteristic function of the interval $\left[\frac{s}{n+1}, \frac{s+1}{n+1}\right]$ w.r.t $[0,1]$.

Lemma 2.5 *Let $h \in C[0,1]$ and α be a fixed number then, we have*

$$\left\|\hat{K}_{n,\alpha}(h(t);x)\right\| \leq \|h\|.$$

Proof From definition of (1.2), we get

$$\begin{aligned}\left|\hat{K}_{n,\alpha}\left(h(t);x\right)\right| &= \left|(n+1)\sum_{s=0}^{n} p_{n,s}^{(\alpha)}(x)\int_{\frac{s}{n+1}}^{\frac{s+1}{n+1}} h(t)dt\right| \\ &\leq (n+1)\sum_{s=0}^{n} p_{n,s}^{(\alpha)}(x)\int_{\frac{s}{n+1}}^{\frac{s+1}{n+1}} |h(t)|\,dt \\ &\leq \hat{K}_{n,\alpha}\left(1;x\right)\|h\| = \|h\|.\end{aligned}$$

□

3 Direct result

In this section we discuss the rate of convergence for the proposed operators (1.2) in terms of the usual *Lipschitz class* $Lip_M\left(\lambda\right)$.

Let $h \in C[0,1]$, $M \geq 0$ and $0 < \lambda \leq 1$, if $h \in Lip_M(\lambda)$ then the inequality

$$|h(t) - h(x)| \leq M|t-x|^{\lambda}, \quad \forall\, t, x \in [0,1]$$

holds.

Theorem 3.1 *For every $h \in Lip_M(\lambda)$ then, we have*

$$\left|\hat{K}_{n,\alpha}(h(t);x) - h(x)\right| \leq M\mu_n^{\lambda}(x),$$

where

$$\mu_n^2(x) = \frac{1}{3(n+1)^2}\left\{3\left(-n+2\alpha-1\right)x^2 + 3\left(n-2\alpha+1\right)x + 1\right\}.$$

Proof By monotonicity property of the operators $\hat{K}_{n,\alpha}(.;x)$, we get

$$\begin{aligned}\left|\hat{K}_{n,\alpha}(h(t);x) - h(x)\right| &\leq \hat{K}_{n,\alpha}\left(|h(t)-h(x)|;x\right)\\ &\leq (n+1)\sum_{s=0}^{n} p_{n,s}^{(\alpha)}(x)\int_{\frac{s}{n+1}}^{\frac{s+1}{n+1}} |h(t)-h(x)|\,dt\\ &\leq (n+1)\sum_{s=0}^{n} p_{n,s}^{(\alpha)}(x)\int_{\frac{s}{n+1}}^{\frac{s+1}{n+1}} |t-x|^{\lambda}dt.\end{aligned}$$

Applying the Hölder's inequality, we get

$$\begin{aligned}\left|\hat{K}_{n,\alpha}(h(t);x) - h(x)\right| &\leq M\left[(n+1)\sum_{s=0}^{n} p_{n,s}^{(\alpha)}(x)\int_{\frac{s}{n+1}}^{\frac{s+1}{n+1}} |t-x|^2dt\right]^{\frac{\lambda}{2}}\\ &\quad\times\left[(n+1)\sum_{s=0}^{n} p_{n,s}^{(\alpha)}(x)\int_{\frac{s}{n+1}}^{\frac{s+1}{n+1}} dt\right]^{\frac{2-\lambda}{2}}\\ &= M\mu_n^{\lambda}(x).\end{aligned}$$

□

4 Bounded variation

In present section, we estimate the rate of the convergence of operators (1.2) in the class of all absolutely continuous functions h defined on the interval $[0,1]$, having a derivative h' equivalent with a function of bounded variation on $[0,1]$, denoted by $DBV[0,1]$. It may be observed that for $h \in DBV[0,1]$, we may write

$$h(x) = \int_0^x f(t)dt + h(0),$$

where $f(t)$ is a function of bounded variation on $[0,1]$.

Lemma 4.1 *For $0 \leq x \leq 1$ and for sufficiently large n, we have*

(i) *Since* $0 \leq y < x$, *therefore*

$$\beta_n(x, y) = \int_0^y U_{n,\alpha}(x, t)\, dt \leq \frac{C}{4(n+1)(x-y)^2}.$$

(ii) *If* $x < z \leq 1$ *then we get*

$$1 - \beta_n(x, z) = \int_z^1 U_{n,\alpha}(x, t)\, dt \leq \frac{C}{4(n+1)(z-x)^2}.$$

Theorem 4.1 *Let* $h \in DBV[0, 1]$ *then for every* $x \in (0, 1)$ *and sufficiently large n, we have*

$$\begin{aligned}\left|\hat{K}_{n,\alpha}(h; x) - h(x)\right| &\leq \frac{1}{4(n+1)}\left|h'(x+) + h'(x-)\right| + \frac{1}{4}\sqrt{\frac{1}{(n+1)}}\left|h'(x+) - h'(x-)\right| \\ &\quad + \frac{x}{\sqrt{n}} V_{x-\frac{x}{\sqrt{n}}}^{x}(h_x') + \frac{C}{4(n+1)x}\sum_{s=1}^{[\sqrt{n}]} \vee_{x-\frac{x}{k}}^{x}\left(h_x{}'\right). \\ &\quad + \frac{(1-x)}{\sqrt{n}} \overset{x+(1-x)/\sqrt{n}}{\underset{x}{V}}(h')_x + \frac{C}{4(n+1)(1-x)}\sum_{s=1}^{[\sqrt{n}]} \overset{x+(1-x)/k}{\underset{x}{V}}((h')_x),\end{aligned}$$

where $\overset{b}{\underset{a}{V}} h(x)$ *denotes the total variation of h on* $[a, b]$, h_x *is a function given by*

$$h_x(t) = \begin{cases} h(t) - h(x-), & 0 \leq t < x, \\ 0, & t = x, \\ h(t) - h(x+), & x < t < 1. \end{cases}$$

Proof Since $\hat{K}_{n,\alpha}(1; x) = 1$, for all $x \in [0, 1]$, we have

$$\begin{aligned}\hat{K}_{n,\alpha}(h(t); x) - h(x) &= \int_0^1 (h(t) - h(x)) U_{n,\alpha}(x; t)\, dt \\ &= \int_0^1 U_{n,\alpha}(x; t)\left(\int_x^t h'(u) du\right) dt, \end{aligned} \tag{4.1}$$

for $h \in DBV[0, 1]$, we can write

$$\begin{aligned}h'(u) &= \frac{1}{2}\left(h'(x+) + h'(x-)\right) + h_x{}'(u) + \frac{1}{2}\left(h'(x+) - h'(x-)\right)\operatorname{sgn}(u - x) \\ &\quad + \delta_x(u)\left(h'(u) - \frac{1}{2}\left(h'(x+) + h'(x-)\right)\right), \end{aligned} \tag{4.2}$$

where

$$\delta_x(u) = \begin{cases} 1, & u = x, \\ 0, & u \neq x. \end{cases}$$

It is obvious that

$$\int_0^1 \left(\int_x^t \left(h'(u) - \frac{1}{2}\left(h'(x+) + h'(x-)\right)\right)\delta_x(u)\, du\right) U_{n,\alpha}(x, t)\, dt = 0. \tag{4.3}$$

Using (2.1), we have

$$\int_0^1 \left(\int_x^t \frac{1}{2} \left(h'(x+) + h'(x-) \right) du \right) U_{n,\alpha}(x,t)\, dt$$
$$= \frac{1}{2} \left(h'(x+) + h'(x-) \right) \hat{K}_{n,\alpha}((t-x);x). \tag{4.4}$$

Moreover,

$$\int_0^1 \left(\int_x^t \frac{1}{2} \left(h'(x+) - h'(x-) \right) \operatorname{sgn}(u-x)\, du \right) U_{n,\alpha}(x,t)\, dt$$
$$= \int_0^1 \frac{1}{2} \left(h'(x+) - h'(x-) \right) (t-x)\, U_{n,\alpha}(x,t)\, dt$$
$$\leq \frac{1}{2} \left| h'(x+) - h'(x-) \right| \int_0^1 |t-x|\, U_{n,\alpha}(x,t)\, dt$$
$$\leq \frac{1}{2} \left| h'(x+) - h'(x-) \right| \hat{K}_{n,\alpha}(|t-x|;x)$$
$$\leq \tfrac{1}{2} \left| h'(x+) - h'(x-) \right| \left(\hat{K}_{n,\alpha}\left((t-x)^2;x\right) \right)^{1/2}. \tag{4.5}$$

Using Lemma 2.3 and from Eqs. (4.2)–(4.5), we obtain

$$\begin{aligned}
\hat{K}_{n,\alpha}(h;x) - h(x) &\leq \frac{1}{2} \left(h'(x+) + h'(x-) \right) \hat{K}_{n,\alpha}((t-x);x) \\
&\quad + \frac{1}{2} \left| h'(x+) - h'(x-) \right| \left(\hat{K}_{n,\alpha}\left((t-x)^2;x\right) \right)^{1/2} \\
&\quad + \int_0^1 \left(\int_x^t h_x{}'(u)\, du \right) U_{n,\alpha}(x,t)\, dt \\
&\leq \frac{1}{4(n+1)} \left(h'(x+) + h'(x-) \right) + \frac{1}{4} \sqrt{\frac{1}{(n+1)}} \left| h'(x+) - h'(x-) \right| \\
&\quad + \int_0^1 \left(\int_x^t h_x{}'(u)\, du \right) U_{n,\alpha}(x,t)\, dt.
\end{aligned}$$

Therefore

$$\begin{aligned}
\left| \hat{K}_{n,\alpha}(h;x) - h(x) \right| &\leq \frac{1}{4(n+1)} \left| h'(x+) + h'(x-) \right| + \frac{1}{4} \sqrt{\frac{1}{(n+1)}} \left| h'(x+) - h'(x-) \right| \\
&\quad + J_{1,n,x} + J_{2,n,x},
\end{aligned} \tag{4.6}$$

where

$$J_{1,n,x} = \left| \int_0^x \left(\int_x^t h_x{}'(u)\, du \right) U_{n,\alpha}(x,t)\, dt \right|,$$

and

$$J_{2,n,x} = \left| \int_x^1 \left(\int_x^t h_x{}'(u)\, du \right) U_{n,\alpha}(x,t)\, dt \right|.$$

Applying Lemma 2.5, by integrating by parts and take $y = x - \frac{x}{\sqrt{n}}$, we have

$$\begin{aligned} J_{1,n,x} &= \left| \int_0^x \left(\int_x^t h_x{}'(u)\, du \right) U_{n,\alpha}(x,t)\, dt \right| \\ &= \left| \int_0^x \beta_n(x,t) h_x^{'}(t) dt \right| \\ &\leq \int_0^y \beta_n(x,t) \left| h_x^{'}(t) \right| dt + \int_y^x \beta_n(x,t) \left| h_x^{'}(t) \right| dt \end{aligned}$$

Since $h_x^{'}(x) = 0$ and $\beta_n(x,t) \leq 1$, it implies

$$\begin{aligned} \int_{x-\frac{x}{\sqrt{n}}}^x \beta_n(x,t) \left| h_x^{'}(t) \right| dt &= \int_{x-\frac{x}{\sqrt{n}}}^x \left| h_x^{'}(t) - h_x^{'}(x) \right| \beta_n(x,t) dt \\ &\leq \int_{x-\frac{x}{\sqrt{n}}}^x V_t^x(h_x^{'}) dt \leq \frac{x}{\sqrt{n}} V_{x-\frac{x}{\sqrt{n}}}^x(h_x^{'}). \end{aligned}$$

Again using Lemma 2.5 and put $y = x - \frac{x}{\sqrt{n}}$,

$$\begin{aligned} \int_0^{x-\frac{x}{\sqrt{n}}} \left| h_x{}'(t) \right| \beta_n(x,t)\, dt &\leq \frac{C}{4(n+1)} \int_0^{x-\frac{x}{\sqrt{n}}} \frac{\left| h_x{}'(t) \right|}{(x-t)^2} dt \\ &\leq \frac{C}{4(n+1)x} \int_1^{\sqrt{n}} \vee_{x-\frac{x}{u}}^x \left(h_x{}' \right) du \\ &\leq \frac{C}{4(n+1)x} \sum_{s=1}^{[\sqrt{n}]} \vee_{x-\frac{x}{k}}^x \left(h_x{}' \right). \end{aligned}$$

Thus

$$J_{1,n,x} \leq \frac{x}{\sqrt{n}} V_{x-\frac{x}{\sqrt{n}}}^x(h_x^{'}) + \frac{C}{4(n+1)x} \sum_{s=1}^{[\sqrt{n}]} \vee_{x-\frac{x}{k}}^x \left(h_x{}' \right). \tag{4.7}$$

By using Lemma 2.5 we can write

$$J_{2,n,x} \leq J_{2,n,x}^{(1)} + J_{2,n,x}^{(2)},$$

where

$$J_{2,n,x}^{(1)} = \left| \int_x^z \left(\int_t^x h_x^{'}(u) du \right) d_t \left(1 - \beta_n(x,t) \right) \right|,$$

and

$$J_{2,n,x}^{(2)} = \left| \int_z^1 \left(\int_t^x h_x^{'}(u) du \right) d_t \left(1 - \beta_n(x,t) \right) \right|.$$

Applying integration by parts as well as using Lemma 2.5, since $(1-\beta_n(x,t)) \leq 1$ and putting $z = x + \frac{(1-x)}{\sqrt{n}}$, we have

$$\begin{aligned} J_{2,n,x}^{(2)} &= \left| \int_z^1 \left(\int_t^x h_x^{'}(u)du \right) d_t \, (1-\beta_n(x,t)) \right| \\ &\leq \frac{C}{4(n+1)} \int_z^1 \overset{t}{\underset{x}{V}}(h')_x \frac{1}{(t-x)^2} dt \\ &\leq \frac{C}{4(n+1)} \int_{x+\frac{(1-x)}{\sqrt{n}}}^1 \overset{t}{\underset{x}{V}}(h')_x \frac{1}{(t-x)^2} dt. \end{aligned}$$

By substituting $u = (1-x)/(t-x)$, we obtain

$$\begin{aligned} J_{2,n,x}^{(2)} &\leq \frac{C}{4(n+1)(1-x)} \int_1^{\sqrt{n}} \overset{x+(1-x)/u}{\underset{x}{V}}(h')_x du \\ &\leq \frac{C}{4(n+1)(1-x)} \sum_{k=1}^{[\sqrt{n}]} \int_k^{k+1} \overset{x+(1-x)/k}{\underset{x}{V}}(h')_x du \\ &\leq \frac{C}{4(n+1)(1-x)} \sum_{s=1}^{[\sqrt{n}]} \overset{x+(1-x)/s}{\underset{x}{V}}((h')_x). \end{aligned}$$

Again applying integration by parts as well as using Lemma 2.5, for calculating $J_{2,n,x}^{(1)}$, since $(1-\beta_n(x,t)) \leq 1$ and putting $z = x + \frac{(1-x)}{\sqrt{n}}$, we get

$$\begin{aligned} J_{2,n,x}^{(1)} &= \left| \int_x^z \left(\int_t^x h_x^{'}(u)du \right) d_t \, (1-\beta_n(x,t)) \right| \leq \int_x^z \overset{t}{\underset{x}{V}}(h')_x dt \\ &\leq \frac{(1-x)}{\sqrt{n}} \overset{x+(1-x)/\sqrt{n}}{\underset{x}{V}}(h')_x = J_{2,n,x}^{(1)} + J_{2,n,x}^{(2)}. \end{aligned}$$

Now

$$J_{2,n,x} \leq \frac{(1-x)}{\sqrt{n}} \overset{x+(1-x)/\sqrt{n}}{\underset{x}{V}}(h')_x + \frac{C}{4(n+1)(1-x)} \sum_{s=1}^{[\sqrt{n}]} \overset{x+(1-x)/s}{\underset{x}{V}}((h')_x) \tag{4.8}$$

From Eqs. (4.6)–(4.8) then we have

$$\left|\hat{K}_{n,\alpha}\left(h;x\right)-h\left(x\right)\right| \leq \frac{1}{4(n+1)}\left|h'\left(x+\right)+h'\left(x-\right)\right| + \frac{1}{4}\sqrt{\frac{C}{(n+1)}}\left|h'\left(x+\right)-h'\left(x-\right)\right|$$
$$+\frac{x}{\sqrt{n}}V_{x-\frac{x}{\sqrt{n}}}^{x}\left(h_x^{'}\right)+\frac{C}{4(n+1)x}\sum_{s=1}^{[\sqrt{n}]}\vee_{x-\frac{x}{s}}^{x}\left(h_x{}'\right).$$
$$+\frac{(1-x)}{\sqrt{n}}\underset{x}{\overset{x+(1-x)/\sqrt{n}}{V}}\left(h'\right)_x+\frac{C}{4(n+1)(1-x)}\sum_{s=1}^{[\sqrt{n}]}\underset{x}{\overset{x+(1-x)/s}{V}}\left(\left(h'\right)_x\right).$$

The Theorem is proved. □

5 Voronovskaya type theorem

In this section we prove a Voronovskaya type result for the α-Bernstein–Kantorovich operators.

Theorem 5.1 *Let $h \in C[0,1]$. If h is twice differentiable in $x \in [0,1]$ and h'' is continuous at x, then the following limit holds:*

$$\lim_{n\to\infty} n\left|\hat{K}_{n,\alpha}(h(t);x)-h(x)\right| \leq \tfrac{(1-2x)}{2}h'(x)+\tfrac{(1-x)x}{2}h''(x)\,.$$

Proof Using Taylor's formula there exists η lying between x and t such that

$$h(t)=h(x)+(t-x)h'(x)+\frac{(t-x)^2}{2!}h''(x)+r(t,x)(t-x)^2,$$

where

$$r\left(t,x\right)=\frac{h''\left(\eta\right)-h''\left(x\right)}{2},$$

and h is a continuous function which vanishes as $t\to x$. Applying the operator $\hat{K}_{n,\alpha}$ to the above equality, we get

$$\hat{K}_{n,\alpha}\left(h(t)-h(x);x\right)=h'(x)\hat{K}_{n,\alpha}\left((t-x);x\right)+\frac{h''(x)}{2!}\hat{K}_{n,\alpha}\left((t-x)^2;x\right)$$
$$+\hat{K}_{n,\alpha}\left(r(t,x)(t-x)^2;x\right).$$

Therefore

$$\lim_{n\to\infty} n\left[\hat{K}_{n,\alpha}\left(h(t)-h(x);x\right)\right]=h'(x)\lim_{n\to\infty} n\left[\hat{K}_{n,\alpha}\left((t-x);x\right)\right]$$
$$+\frac{h''(x)}{2!}\lim_{n\to\infty}\left[\hat{K}_{n,\alpha}\left((t-x)^2;x\right)\right]$$
$$+\lim_{n\to\infty} n\left[\hat{K}_{n,\alpha}\left(r(t,x)(t-x)^2;x\right)\right]. \qquad (5.1)$$

In order to estimate the last term of Eq. (5.1), for every $\varepsilon>0$ choose $\delta>0$ such that

$$r(t,x)<\varepsilon \quad \text{for } |t-x|<\delta.$$

Therefore if $|t-x| < \delta$ then $\left|r(t,x)(t-x)^2\right| < \varepsilon(t-x)^2$ while if $|t-x| \geq \delta$, then since $r(t,x) < M$ we have

$$\left|r(t,x)(t-x)^2\right| \leq \frac{M}{\delta^2}(t-x)^4.$$

So we can write

$$\hat{K}_{n,\alpha}\left(r(t,x)(t-x)^2;x\right) \leq \varepsilon\mu_n^2 + \frac{M}{\delta^2}\mu_n^4.$$

By direct calculations we obtain

$$\mu_n^4(x) = O\left(\frac{1}{n^2}\right),$$

and we conclude

$$\lim_{n\to\infty} n\left[\hat{K}_{n,\alpha}\left(r(t,x)(t-x)^2;x\right)\right] = 0.$$

We achieve the required result. □

Acknowledgements The authors are extremely thankful to the editor and referee for valuable comments, leading to a better presentation of the article.

References

1. Acar, T., Acu, A.M., Manav, N.: Approximation of functions by genuine Bernstein–Durrmeyer type operators. J. Math. Ineqal. **12**(4), 975–987 (2018)
2. Chen, X., Tan, J., Liu, Z., Jin, Xie J.: Approximation of functions by a new family of generalized Bernstein operators. J. Math. Anal. Appl. **450**, 244–261 (2017)
3. Deo, N., Dhamija, M.: Better approximation results by Bernstein–Kantorovich operators. Lobachevskii J. Math. **38**(1), 94–100 (2017)
4. Dhamija, M., Deo, N.: Approximation by generalized positive linear-Kantorovich operators. Filomat **31**(14), 4353–4368 (2017)
5. Dhamija, M., Pratap, R., Deo, N.: Approximation by Kantorovich form of modified Szász–Mirakyan operators. Appl. Math. Comput. **317**, 109–120 (2018)
6. Gonska, H., Heilmann, M., Raşa, I.: Kantorovich operators of order k. Numer. Funct. Anal. Optim. **32**(7), 717–738 (2011)
7. Kajla, A., Araci, S.: Blending type approximation by Stancu-Kantorovich operators based on Pólya Eggenberger distribution. Open Phys. **15**, 335–343 (2017)
8. Li, B., Zhou, D.: Analysis of approximation by linear operators on variable $L_\rho^{p(.)}$ spaces and applications in learning theory. Abstr. Appl. Anal. **2014**, 1–10 (2014). https://doi.org/10.1155/2014/454375
9. Mohiuddine, S.A., Acar, T., Alotaibi, A.: Construction of a new family of Bernstein-Kantorovich operators. Math. Methods Appl. Sci. **40**(18), 7749–7759 (2017)
10. Özarslan, M.A., Duman, O.: Smoothness poperties of modified Bernstein–Kantorovich operators. Numer. Funct. Anal. Optim. **37**(1), 92–105 (2016)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2019.2954779, IEEE Transactions on Information Forensics and Security

Adaptive Weighted Graph Approach to Generate Multimodal Cancelable Biometric Templates

Gurjit Singh Walia, *Senior Member, IEEE*, Gaurav Jain, Nipun Bansal, Kuldeep Singh, *Senior Member, IEEE*

***Abstract*— Multimodal biometric systems offer numerous advantages over unimodal counterparts and are being used extensively in diverse applications. However, fusion of biometric data is a non-trivial task and curtail employability of multimodal systems for a varying set of biometric characteristics with different type and dimension. Moreover, comprehensive solutions against adversary attacks that ensure template protection and prevent presentation attacks are not in place. In this paper, a secure multimodal cancelable biometric system is proposed to address these concerns. This approach introduces key images based generic feature extraction technique which reduces feature dimension and achieves revocability. The non-invertibility and unlinkability are ensured through cross-diffusion of complementary information from different modalities. A new feature fusion method based on an adaptive graph is proposed to generate multimodal cancelable biometric templates. Robustness against presentation attack is accomplished through quality based adaptation of features. Extensive experimentation is performed on benchmark databases for fingerprint, face, and iris, to illustrate the efficacy of multimodal cancelable templates. The proposed approach is shown to perform favorably against state-of-the-art feature fusion methods. Furthermore, the resilience of the proposed approach against security and privacy attacks is demonstrated.**

***Index Terms*—Multimodal, Cancelable, Adaptive, Fusion**

I. Introduction

BIOMETRIC recognition systems are being extensively employed for security-critical applications. The rapid proliferation of biometric systems and their application in cloud environments has led to various concerns regarding system security and potential privacy breaches. These systems store biometric data that can be vulnerable not only to adversary attacks via public network but also to elicit disclosure under captive environments. Biometric data is considered as a piece of sensitive information because unlike passwords, biometric characteristics can neither be reissued nor be revoked. Compromise of biometric data results in permanent loss of an individuals identity, and hence, is a growing concern in today's digital world. Biometric systems can be compromised mainly due to issues, which can be loosely classified as: (1) *template database leakage*, and (2) *presentation attacks* (e.g., using gummy fingers, face mask, dismembered fingers from a legitimate user) [1]. While presentation attacks can induce illegitimate access, template leakage may risk in covert hearing and unauthorized access to private information. Hence, these vulnerabilities may incapacitate the application of biometric systems under cloud environments via public networks.

Gurjit Singh Walia is with the SAG, Defence Research and Development Organisation, New Delhi, India. (e-mail: gurjit.walia@gmail.com)

Gaurav Jain and Nipun Bansal are with the Department of Computer Science & Engineering, Delhi Technological University, New Delhi, India.

Kuldeep Singh is with the Department of Electronics & Communication Engineering, Malviya National Institute of Technology, Jaipur, India

To protect biometric data, the concept of *cancelable biometrics* was introduced by Ratha et al. [2]. It involves an intentional transformation of original biometric templates such that enrolment and verification are performed in the transformed domain. Transformed templates should be robust to adversary attacks and hence must satisfy the following properties, (1) *Revocability*, (2) *Unlinkability*, (3) *Non-invertibility*, while preserving the (4) *Performance* [3]. Generally, cancelable biometric approaches can be classified into (1) *Biometric salting*, and (2) *Non-invertible transforms*. In biometric salting, the biometric template is mixed with an artificial pattern which may be random noise or synthetic pattern. The dilemma in deciding the relative magnitude of the noise to be added makes it difficult to employ salting techniques. Besides, if the artificial pattern is stolen, the original biometric characteristic can be extracted [4]. In contrast, non-invertible transformation based cancelable approaches mutate the original biometric templates such that it is infeasible to invert the transformed template. Predominantly used non-invertible transformation techniques are *Random Projections* [5], *Bloom Filters* [6], and *Biohashing* [7]. However, most of the non-invertible transform-based cancelable approaches were demonstrated for unimodal biometric systems [8], [9]. In unimodal systems, the trade-off between security and performance limits the success of cancelable approaches. Besides, unimodal systems suffer from the inevitable issues of non-universality, noisy data, inter-class variation, and presentation attacks.

Recently, cancelable biometrics has been extended from unimodal systems to multimodal systems [10]. Generally, multimodal systems are employed to compensate for the loss in accuracy and prevent presentation attacks [11]. Typically, multimodal schemes can be classified on the basis of the kind of information being fused such as *feature level* [6], *decision level* [11], and *score level* [12]. Being more discriminative, feature level fusion is preferred over score or decision level fusion [11]. However, feature level fusion is limitedly investigated due to different type and dimension of captured biometric samples extracted from multiple sensors. Additionally, presentation attacks were avoided through liveness detection which mainly exploits the image quality for its realisation [13]–[16]. However, most of the liveness detection approaches do not address template protection and adaptive fusion of multiple modalities. In sum, exhaustive solutions that cater to adversary attacks, and also maintains high performance are limitedly investigated. Our method is a comprehensive solution for multimodal biometric recognition that not only provides complete non-invertibility of the template but also drastically reduces feature dimension. Proposed key-based weighted graph fusion can be adapted to any modality irrespective of its dimension and type.

The rest of the paper is organized as follows: In Section II,

1556-6013 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2019.2954779, IEEE Transactions on Information Forensics and Security

closely related literature work is reviewed. In Section III, limitations of the existing works, and our contributions are highlighted. In Section IV we present the proposed multibiometric recognition system along with the cancelable template generation approach. Further, we perform experimental validation in Section V. Privacy and security of the proposed method are analyzed in Section VI. Finally, concluding remarks and future directions are drawn in Section VII.

II. Related Work

Biometric security has been rigorously investigated over the past decades. Readers are advised to refer the papers [3] [4] for detailed solutions for biometric security and privacy preservances. In this section, we categorically review the closely related literature mainly considering non-invertible transform for cancelable feature generation and also, countermeasures against presentation attacks.

Rathgeb et al. [17] introduced a bloom filter based non-invertible transform for template protection. In this approach, biometric features were mapped to an irreversible representation i.e. bloom filters. Further, unlinkability was achieved using an application-specific token. But, it was later proved by Bringer et al. [18] that bloom filter was susceptible to cross-matching attacks. To overcome this issue, Sadhya et al. [19] proposed a modified approach on bloom filter, where application-specific token T, was replaced with a key matrix κ. However, this approach performed comparison in the original domain instead of the transformed domain and hence, it was prone to security concerns. Recently, Gomez et al. [6] proposed a generic approach to generate cancelable biometric templates based on bloom filters. In this, the value of keys was computed through statistical analysis of the key features. Mostly, bloom filter based approaches for template protection fail to address the issues of unlinkability and cross-matching attacks [6] [18].

Besides bloom filters, random projections have also been explored to generate cancelable biometric templates. Recently, Kaur and Khanna [20] proposed a template transformation technique dubbed as Random Distance Method (RDM) for generating dimensionally reduced templates. However, non-invertibility was not an inherent property of RDM and was achieved by applying median filtering over the transformed templates. This work on RDM was further extended to address the security issues in cloud-based remote biometric authentication scenarios [21]. On similar lines, Qui et al. [22] experimentally validated that a certain amount of added noise can induce cancelable properties while still meeting the accuracy standards. The authors generated cancelable palmprint templates through random comparison and noise data. Further, hashing based approaches extend the random projection-based approaches by exploiting the properties of hash functions. For instance, Dwivedi et al. [23] employed a look-up table to generate cancelable iris templates. The look-up table was used to map a decimal encoded vector which was generated from the binary features. However, using a look-up table makes the approach vulnerable to inversion attacks, if the transformed features and look-up table are compromised. Sadhya et al. [24], proposed a locality sensitive hashing based approach, dubbed as locality sampled code (LSC), to generate cancelable IrisCodes features. Despite generating a secure template with good recognition performance, this technique was limited to unimodal biometric systems having binary feature representations. Lai et al. [8] introduced Indexing-First-One (IFO) hashing to transform real-valued features into hashed code to generate cancelable iris templates. Similarly, Jin et al. [9] generated fingerprint cancelable templates using "Index-of-Max" (IoM) hashing which was based on a locality sensitive hashing. Although IoM and IFO display good recognition performance along with non-invertibility, application of these techniques was restricted to binary feature representations. In [25], the authors proposed to design a cancelable fingerprint template using an alignment-free minutia descriptor, namely Partial Local Structure (PLS) descriptor. Non-invertible transforms were applied over the PLS descriptor rather than the minutia descriptor using Permutated Randomized Non-Negative Least Square (PR-NNLS) to achieve the desired cancelable properties while maintaining performance. Again, the proposed approach is limited to only fingerprint related applications. Hence, an adaptation of these methods to any type and dimension of feature, along with countermeasure against presentation attacks remains a challenge.

Recently, multibiometric cancelable systems were realized to simultaneously protect biometric data and defeat presentation attacks. Multimodal systems were found to perform better than unimodal counterparts, mainly due to the presence of multiple discriminative sources of information which are difficult to spoof simultaneously [11]. For instance, Sadhya et al. [26] fused soft biometric characteristics with primary biometric characteristics using a bayesian approach. However, in this approach performance degrades if the comparator's characteristics were not known. Also, soft biometric characteristics exhibit low discriminatory information. W. Yang et al. [27], proposed to fuse fingerprint and finger-vein using a non-invertible Enhanced Partial Discrete Fourier Transform (EP-DFT). In [12], the authors proposed to generate cancelable multibiometric templates using graph-based random walk cross-diffusion and optimal score fusion. Bio-hashing based data type conversion was applied to fuse fingerprint minutiae and finger-vein feature. Zhong et al. [28] proposed to combine the accuracy of palmprint and liveness detection capabilities of dorsal hand veins (DHV) to generate multibiometric cancelable templates. For each biometric characteristic, 128-bit binary encoding was generated using a deep hashing network (DHN), along with with a biometric graph matching (BGM) approach for effective biometric verification. Mostly, feature level fusion techniques are limited to a particular set of modalities, and cannot be extended to other feature representations. Hence, it is observed that compared to other levels of fusion, the potential of feature-level fusion has not been exploited as much. The complexity associated with the application of multibiometric systems led researchers to explore alternative solutions to prevent presentation attacks [29]. Mainly, liveness detection approaches were employed to address this issue [13]. In [14] image quality parameters like texture, contrast, brightness helped identify masqueraders

1556-6013 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2019.2954779, IEEE Transactions on Information Forensics and Security

and prevent presentation attacks. In [15], Fernandez et. al. proposed a quality-based multibiometric system. In this, interoperability of this system was improved by conditionally fusing comparison scores, based on image quality. A similar approach was described by Poh et. al. [16], which employed a quality-based score normalization technique for multimodal biometric fusion.

For security-critical applications, an adversary may not only camouflage as a legitimate user but also try to illicitly steal biometric data. Hence, comprehensive security against adversary attacks must consider both template security and prevention from presentation attacks.

III. Motivations and Contributions

Investigating the existing biometric security mechanisms revealed the following limitations:

- *Non-invertibility:* most of the non-invertible transformation approaches tend to be vulnerable to partial or complete inversion (e.g., [30] [7]). In the scheme proposed in [30], multiple swindled templates may result in inversion to obtain the original biometric characteristic. Achieving complete non-invertibility is particularly challenging because increased non-invertibility results in reduced template distinctiveness, which degrades the performance.
- *Generality:* cancelable template generation is often limited to a certain class of feature extraction methods. For instance, the application of hashing-based approach proposed by Lin et. al. [9] was restricted to modalities that can be represented as binary features and have a fixed length. Similarly, in [20], random projection approach was restricted to features which have a fixed length represented as a one-dimensional vector. Also, the fusion of non-homogeneous feature representations is again a challenging task.
- *Image Quality Conundrum:* performance of most of the biometric systems is dependent on image quality. Poor image quality results in low-quality features and hence low recognition rates. However, this enables biometric systems to detect presentation attacks. Whereas, if the system is quality independent, presentation attack goes unnoticed but persistent performance levels are achieved.
- *Robustness to Adversary Attacks:* most of the cancelable biometric approaches concern with only one class of adversary attacks i.e., template security [9]. On the other hand, liveness detection and presentation attacks were dealt with separately [13]. Generally, the comprehensive solution against adversary attacks was not considered.

To overcome these limitations, we propose an adaptive weighted graph approach to generate multimodal cancelable templates. The main contributions of our work are as follows:

- We propose a general framework to generate multimodal cancelable biometric templates. This is achieved through the novel feature set computation algorithm, which generates similarity graphs using key images. The proposed framework is deemed as a universal framework which is independent of the biometric characteristic used, or the feature extraction method employed. In addition, the generated feature set is dimensionally reduced, improving speed and reducing space requirements.
- We introduce a nonlinear graph fusion approach which fuses complementary information from different biometric characteristics. The nonlinear graph fusion approach has multiple advantages. First, strong information from different modalities is captured, while any weak information is suppressed. Second, nonlinearity enables the fused templates to achieve non-invertibility.
- We put forward an adaptive approach to resolve the *"image quality conundrum"*. A no-reference image quality analysis is employed to weigh features from respective modalities. This enables fused templates to adapt to features having superior image quality and suppress features with low image quality. Through this, the proposed biometric system can differentiate between 'presentation attacks' and 'mere low-quality images'.
- We rigorously analyze the robustness against security and privacy concerns. In particular, non-invertibility, unlinkability of the generated templates are tested. Further, robustness against various attacks namely false accept attacks, brute-force attacks, ARM attacks, and substitution attacks is also highlighted.

The following section describes the proposed cancelable multibiometric framework in detail.

IV. Proposed Methodology

The proposed multimodal biometric system leverages key images based cancelable feature generation, along with its quality based adaptation for robust recognition in an adversarial environment. The system is realized to perform recognition using commonly used biometric characteristics, namely fingerprint (p), face (f), and iris (i). Fig. 1 illustrates the overview of the proposed multibiometric system. Generic feature extraction for each modality is performed through key images, which can be revoked on the compromise of biometric data. For this, input multibiometric data is captured for three modalities to extract raw features $\delta^{(p)}, \delta^{(f)}$ and $\delta^{(i)}$ respectively. Also, raw features for key images $K_j^{(k)}$ $\forall j \in \{1, 2, ..., n\}, k \in \{p, f, i\}$ are extracted for three modalities as $\psi^{(p)}, \psi^{(f)}, \psi^{(i)}$ respectively. Here, a set of n key images were used for each modality, and these keys were common to all input and query images for all subjects. Association of query images $I^{(k)} \in \{I^{(p)}, I^{(f)}, I^{(i)}\}$ with the corresponding key images $K_j^{(k)}$ $\forall j \in \{1, 2, ..., n\}$ is determined to construct similarity graphs $\mathbf{G}^{(k)} \in \{\mathbf{G}^{(p)}, \mathbf{G}^{(f)}, \mathbf{G}^{(i)}\}$. Further, anchored normalisation is applied over the individual similarity graphs to generate corresponding normalised graphs, $\mathbf{Q}^{(k)} \in \{\mathbf{Q}^{(p)}, \mathbf{Q}^{(f)}, \mathbf{Q}^{(i)}\}$. These normalised graphs are fused using the proposed *Adaptive Weighted Graph Fusion* (AWGF), which not only extracts complementary information from different modalities, but also adaptively weighs each modality in concurrence with image quality. For this, characteristic information of each modality is extracted as sparse graphs $\{\mathbf{S}^{(p)}, \mathbf{S}^{(f)}, \mathbf{S}^{(i)}\}$ and rank graphs $\{\mathcal{R}^{(p)}, \mathcal{R}^{(f)}, \mathcal{R}^{(i)}\}$. These graphs are diffused using proposed nonlinear graph fusion to generate fused vectors $\mu^{(\mathbf{k})} \in \{\mu^{(p)}, \mu^{(f)}, \mu^{(i)}\}$ for respective modalities. Fused vectors are

1556-6013 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2019.2954779, IEEE Transactions on Information Forensics and Security

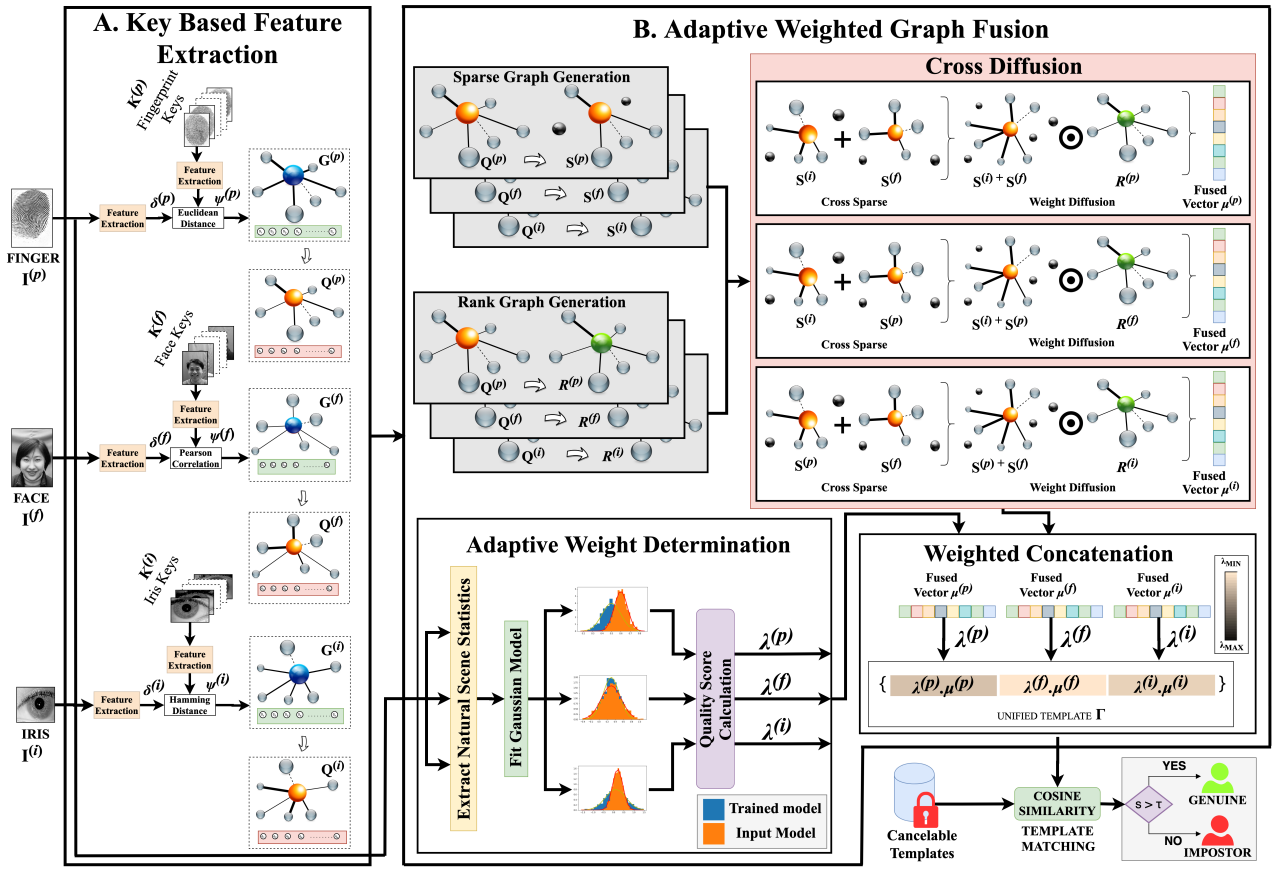


Fig. 1. Overview of the proposed multibiometric system. Normalized graphs constructed through key-based feature extraction is subjected to the proposed cross-diffusion to generate fused vectors. These fused vectors attain weights, determined adaptively using image quality, to generate a cancelable template.

assigned weights $\{\lambda^{(p)}, \lambda^{(f)}, \lambda^{(i)}\}$ adaptively by evaluating quality-metrics for each modality. A weighted concatenation results in the unified feature vector $\mathbf{\Gamma}_q$, which is the robust cancelable biometric template. The generated biometric template is compared using cosine similarity with stored database of templates for making decision about query subject. The proposed approach is elaborated in the following subsections.

A. Key Based Generic Feature Extraction

Generally, feature extraction from multiple modalities generates features with different type and dimension. This limits the applicability of feature fusion methods to only certain physiological characteristics. To overcome this, a key image based generic feature extraction is introduced, which not only achieves the generic nature of the proposal with reduced feature dimension but also ensures high revocability. In the case of template leakage, biometric templates can be regenerated either by changing the key order or by using a new set of key images for each biometric characteristic. Details about generic feature computation are as follows:

1) Graph Generation: Generic nature of the proposal is achieved through the establishment of an association of query images with a fixed set of key images, corresponding to each modality. This leads to construction of non-linear graphs $\mathbf{G}^{(k)} \in \{\mathbf{G}^{(p)}, \mathbf{G}^{(f)}, \mathbf{G}^{(i)}\}$, having a fixed number of nodes n, irrespective of the modality.

For fingerprint, raw features extracted through minutiae-based approach [31] are subjected to fingerprint keys association for the generation of a nonlinear graph $\mathbf{G^{(p)}}$. To determine minutiae location, 3×3 sliding window (centred at pixel u) is traversed in a circularly anticlockwise manner, to generate the *crossing number* (CN):

$$\mathrm{CN}(u) = \frac{1}{2}\sum_{i=1}^{8} |\S(u_{(i \bmod 8)}) - \S(u_{i-1})| \qquad (1)$$

where $\S(u_i) \in \{0, 1\}$, is the pixel intensity of the i^{th} block of the window, centred at pixel u. Each minutiae is represented as vector $M = [x, y, \mathrm{CN}, \theta, \mathit{flag}]$, where (x, y) represent the coordinates of a pixel u, θ defines the orientation of the minutiae, and $\mathit{flag} \in \{0, 1\}$. Hence, for query image $I^{(p)}$, raw feature $\delta^{(p)} \in \mathbb{R}^{m \times 5}$ containing m minutiae is formed such that $\delta^{(p)} = [M_1, M_2, ..., M_m]$. Similarly, raw features $\psi_j^{(p)}, \forall j \in \{1, 2, ..., n\}$ are extracted for fingerprint keys $K_j^{(p)}$. These extracted raw features lead to the formation of nonlinear graph $\mathbf{G}^{(p)} = \{V^{(p)}, E^{(p)}, \mathbf{W}^{(p)}\}$. Wherein, $n+1$ nodes, i.e. $V^{(p)} = \{\delta^{(p)}, \psi_1^{(p)}, \psi_2^{(p)} ... \psi_n^{(p)}\}$ depicts the raw features for query $\delta^{(p)}$ and for fingerprint keys $\{\psi_1^{(p)}, ..., \psi_n^{(p)}\}$. Graph edges $E^{(p)} = \{(\delta^{(p)}, \psi_j^{(p)}) | \forall j \in \{1, 2, ..., n\}\}$ represents similarity of query image with the fingerprint keys. Hence,

1556-6013 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2019.2954779, IEEE Transactions on Information Forensics and Security

each edge is assigned weight $\mathbf{W}_j^{(p)}$, which is defined as:

$$\mathbf{W}_j^{(p)}\left(\delta^{(p)},\psi_j^{(p)}\right)=\frac{n_{match}^2}{n_{\delta^{(p)}}\ n_{\psi_j^{(p)}}},\ \forall j\in\{1,2,...,n\} \quad (2)$$

where $n_{\delta^{(p)}}$ and $n_{\psi_j^{(p)}}$ represent the total number of minutiae extracted from the query and j^{th} fingerprint key respectively, and n_{match} represents the number of minutiae that satisfy the error constraints for coordinates (x,y) and orientation (θ), defined as:

$$\begin{aligned} &d(M_{\delta_m},M_{\psi_m})=\sqrt{(x_{\delta_m}-x_{\psi_m})^2+(y_{\delta_m}-y_{\psi_m})^2}<d_\alpha\\ &d(M_{\delta_m},M_{\psi_m})=\min(|\theta_{\delta_m}-\theta_{\psi_m}|,\ 360^\cdot-|\theta_{\delta_m}-\theta_{\psi_m}|)<d_\theta \end{aligned} \quad (3)$$

For second modality, i.e. face, raw features are extracted using 2D Gabor filters [32], where 2D Gabor filter is defined as:

$$\begin{aligned} &G(x,y)=\frac{\eta^2}{\pi\alpha\beta}\exp\left(-\frac{p^2+\alpha^2q^2}{2\sigma^2}\right)\exp\left(i2\pi\eta p+\phi\right)\\ &p=x\cos\theta+y\sin\theta\ ;\ q=-x\sin\theta+y\cos\theta \end{aligned} \quad (4)$$

where η represents the sinusoidal frequency, ϕ is the phase offset, θ is the orientation, σ is the standard deviation, and α is the spatial aspect ratio. Query image $I^{(f)}$, is convolved with 40 Gabor filters (5 scales, 8 orientations) to generate feature images. These feature images are then reshaped and concatenated to form a 1-D raw feature vector $\delta^{(f)}$. Similarly, each face keys $K_j^{(f)}$ is convolved with the Gabor filter bank to generate raw feature vector $\psi_j^{(f)}$, $\forall j\in\{1,2,...,n\}$.

These extracted raw feature vectors are exploited for the formation of nonlinear graph $\mathbf{G}^{(f)}=\{V^{(f)},E^{(f)},\mathbf{W}^{(f)}\}$. Nodes of the graph $V^{(f)}=\{\delta^{(f)},\psi_1^{(f)},\psi_2^{(f)}...\psi_n^{(f)}\}$ are obtained using query image feature $\delta^{(f)}$, and n face keys features $\psi_j^{(f)}$, $\forall j\in\{1,2,...,n\}$. Graph edges $E^{(f)}=\{(\delta^{(f)},\psi_j^{(f)})|\forall j\in\{1,2,...,n\}\}$ depict the association among the nodes. Hence, similarity weight matrix $\mathbf{W}^{(f)}=\{\mathbf{W}_1^{(f)},\mathbf{W}_2^{(f)},...,\mathbf{W}_n^{(f)}\}$ is determined, where j^{th} edge weight $\mathbf{W}_j^{(f)}$ is defined as:

$$\mathbf{W}_j^{(f)}\left(\delta^{(f)},\psi_j^{(f)}\right)=\mathrm{cov}(\delta^{(f)},\psi_j^{(f)})/(\sigma_{\delta^{(f)}}.\sigma_{\psi_j^{(f)}}) \quad (5)$$

For iris trait, we employed 1D Log-Gabor filter method [33] to extract raw binary features. For this, segmented iris image is normalised using Daugmans rubber sheet model. Normalised iris image is convolved with 1D Log Gabor wavelets with frequency response defined as:

$$G(\eta)=\exp\left(\frac{-(\log(\eta/\eta_o))^2}{2(\log(\sigma/\eta_o))^2}\right) \quad (6)$$

where η_o is the centre frequency and σ controls the bandwidth. The extracted phase data is quantised to four levels. This results in a unique binary pattern, forming the raw feature template $\delta^{(i)}$ corresponding to iris query. Further, raw feature $\psi_j^{(i)}$ is generated for each iris key $K_j^{(i)},\forall j\in\{1,2,...,n\}$.

Similarly, iris graph $\mathbf{G}^{(i)}=\{V^{(i)},E^{(i)},\mathbf{W}^{(i)}\}$ is constructed with $n+1$ nodes. The width of the edges forms a similarity weight matrix $\mathbf{W}^{(i)}$. Elements of similarity weight matrix, $\mathbf{W}_j^{(i)}$ $\forall j\in\{1,2,...,n\}$, Eq. 7, are determined through radial basis function applied over Hamming distance ($\mathbf{H}_\mathrm{d}$).

$$\mathbf{W}_j^{(i)}\left(\delta^{(i)},\psi_j^{(i)}\right)=\exp\left(-\mathbf{H}_\mathrm{d}(\delta^{(i)},\psi_j^{(i)})^2/(2\sigma^2)\right) \quad (7)$$

where σ is scaling factor, and Hamming distance ($\mathbf{H}_\mathrm{d}$), Eq. 8, provides dissimilarity between the iris query and keys features.

$$\mathbf{H}_\mathrm{d}(\delta^{(i)},\psi_j^{(i)})=\frac{1}{N}\sum_{y=1}^{N}\delta_y^{(i)}\oplus\psi_{j_y}^{(i)} \quad (8)$$

In order to preserve individual subject privacy and system security, we have chosen the set of key images, such that $K^{(k)}\notin I^{(k)},\forall k\in\{p,f,i\}$, which leads to indistinguishable generic features of individual modality. On the other hand, generic features of different modality are highly irregular and non-uniform due to differing environment and evaluation approach. Hence, non-linear graphs $\mathbf{G}^{(k)}\in\{\mathbf{G}^{(p)},\mathbf{G}^{(\mathbf{f})},\mathbf{G}^{(i)}\}$ are normalised to achieve distinguishable individual modality generic features. Details of graph normalisation process is discussed as follows.

2) Graph Normalisation: In order to obtain unbiased and distinguishable features, nonlinear graphs $\mathbf{G}^{(\mathbf{k})}$, are normalised using anchored normalisation to generate normalised similarity graphs $\mathbf{Q}^{(k)}\in\{\mathbf{Q}^{(p)},\mathbf{Q}^{(f)},\mathbf{Q}^{(i)}\}$. For this, each non-linear graph $\mathbf{G}^{(k)}$ is normalised using an anchor $A^{(k)}$, which is derived from the impostor score distribution of the k^{th} modality, defined as: $A^{(k)}=A_{avg}^{(k)}+A_{std}^{(k)}$, where $A_{avg}^{(k)}$ and $A_{std}^{(k)}$ refer to the mean and standard deviation of the impostor score distribution of the k^{th} modality. Our approach modifies the approach described in [34], by considering only impostor scores for computing the anchor. To make the scores more distinguishable within a modality and scale the scores of different modalities to the same level, weight transformation, Eq. 9 has been applied to the get normalised graph $\mathbf{Q}^{(k)}\in\{\mathbf{Q}^{(p)},\mathbf{Q}^{(f)},\mathbf{Q}^{(i)}\}$, where the graph weight matrix $\mathcal{Q}_j^{(k)}$, $\forall j\in\{1,2,...,n\}$, $\forall k\in\{p,f,i\}$ are determined as:

$$\mathcal{Q}_j^{(k)}=\begin{cases}\frac{W_j^{(k)}-min\left(W^{(k)}\right)}{2\left(A^{(k)}-min\left(W^{(k)}\right)\right)}, & W_j^{(k)}\le A^{(k)}\\ 0.5+\left(\frac{W_j^{(k)}-A^{(k)}}{max\left(W^{(k)}\right)-A^{(k)}}\right), & W_j^{(k)}>A^{(k)}\end{cases} \quad (9)$$

Algorithm 1 describes the proposed key based generic feature extraction technique. These normalised graphs with distinguishable scores are subjected to proposed fusion approach as discussed in the following subsection.

B. Adaptive Weighted Graph Fusion

Adaptive Weighted Graph Fusion (AWGF) is proposed to achieve complete non-invertibility and robustness to presentation attacks. It comprises of: (1) *Information Mining* to extract complementary information and suppress outliers, (2) *Cross Difussion* to ensure complete non-invertibility, and (3) *Adaptive Unification* to prevent presentation attacks. Details about the proposed AWGF is as follows:

1) Information Mining: To capture complementary information from three modilities, sparse graphs $\mathbf{S}^{(k)}\in\{\mathbf{S}^{(p)},\mathbf{S}^{(f)},\mathbf{S}^{(i)}\}$ and rank graphs $\mathcal{R}^{(k)}\in\{\mathcal{R}^{(p)},\mathcal{R}^{(f)},\mathcal{R}^{(i)}\}$ are constructed from the normalised graphs $\mathbf{Q}^{(k)}\in\{\mathbf{Q}^{(p)},\mathbf{Q}^{(f)},\mathbf{Q}^{(i)}\}$. Sparse graphs ensure robustness to noise and dynamic environments while preserving strong information and suppressing weak information from each modality. In

1556-6013 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2019.2954779, IEEE Transactions on Information Forensics and Security

Algorithm 1 : Key Based Generic Feature Extraction

Input: $I^{(k)}$, $K_j^{(k)}\ \forall\ j \in \{1,2,...,n\}$
Output: $\mathbf{Q}^{(k)} \in \{\mathbf{Q}^{(p)}, \mathbf{Q}^{(f)}, \mathbf{Q}^{(i)}\}$

for $k \in \{p,f,i\}$ **do**
  extract raw features for $I^{(k)}$, denoted as $\delta^{(k)}$
  for $j = 1$ to n **do**
    extract raw features for $K_j^{(k)}$, denoted as $\psi_j^{(k)}$
  end for
end for
for $k \in \{p,f,i\}$ **do**
  for $j = 1$ to n **do**
    construct $\mathbf{G}_j^{(k)}$ using Eq. 2, Eq. 5, Eq. 7
    normalise $\mathbf{G}_j^{(k)}$ to $\mathbf{Q}_j^{(k)}$ using Eq. 9
  end for
end for
return $\mathbf{Q}^{(k)} \in \{\mathbf{Q}^{(p)}, \mathbf{Q}^{(f)}, \mathbf{Q}^{(i)}\}$

our approach, we exploit the sparsity of normalized graphs by selecting those key images, that are strongly correlated to the query. Sparse graph, $\mathbf{S}^{(k)} = \{V^{(k)}, E^{(k)}, \xi^{(k)}\}$ is constructed using k-nearest neighbour (KNN) as follows:

$$\xi_j^{(k)} = \begin{cases} \mathcal{Q}_j^{(k)}, & \left(\mathcal{Q}_j^{(k)} \in \text{KNN}_{\mathbf{Q}^{(k)}} | \kappa\right) \\ 0, & \text{Otherwise} \end{cases} \tag{10}$$

where κ controls the sparsity of the graph, and $\mathcal{Q}^{(k)}$ is the weight matrix of the normalised graph of the k^{th} modality such that $k \in \{p,f,i\}$. Edges corresponding to key images that are similar to the query image are retained, while remaining edges are removed, ensuring robustness to noise.

In order to distinguish significant key image from the insignificant ones, each key image is assigned weights according to its rank. For this, ranks are assigned to each key image based on its similarity with the query image, which is determined by the edge weight between the query and the key image in normalized graph. The weight graph $\mathcal{R}^{(k)} = \{V^{(k)}, E^{(k)}, \mathbf{\Upsilon}^{(k)}\}$ having a weight matrix $\mathbf{\Upsilon}^{(k)}$ defined as:

$$\mathbf{\Upsilon}_j^{(k)} = Rank(\mathcal{Q}_j^{(k)}),\ \forall j \in \{1,2,...,n\} \tag{11}$$

where $Rank$, assigns a rank $r \in \{1,2,...,n\}$ to each key image in concurrence with the relative score.

2) Cross Diffusion: Cross diffusion method is devised to effectively strengthen strong relationships between different modalities while suppressing any noise or weak links. To achieve this, we cross diffuse the distinctive information acquired through sparse graph $S^{(k)}$, and rank graph $\mathcal{R}^{(k)}$, $\forall k \in \{p,f,i\}$. While sparse graphs ensure removal of outliers, rank graphs prevent any bias of modality. Hence, cross diffusion of sparse and rank graph maintain a trade-off between removing insignificant information, without mistakenly missing out any information. Proposal is generic in nature and hence can be extended to any type and dimension of the modality. In addition, diffusion of information from multiple modalities makes it even more difficult for an adversary to regenerate biometric characteristics. Essentially, cross diffusion can be summed up as *ORing* of sparse graphs followed by their *ANDing* with rank graphs. To realise cross diffusion, we perform *cross sparse* (Eq. 12) by adding the sparse graphs of two other modalities to generate $\zeta^{(k^{(t)})}$, where $(k^{(t)})$ is t^{th} element of set k. This is followed by *weight diffusion* (Eq. 13) of the rank graph of the self modality with $\zeta^{(k^{(t)})}$ to generate fused vector $\mu^{(k^{(t)})}$.

$$\zeta^{(k^{(t)})} = \sum_Z \xi^{(Z)},\ \text{where } Z \in \{k\} - \{k^{(t)}\} \tag{12}$$

$$\mu^{(k^{(t)})} = \mathbf{\Upsilon}^{(k^{(t)})} \odot \zeta^{(k^{(t)})},\ \forall t \in \{1,2,3\} \tag{13}$$

To compute $\mu^{(k)}$, key images that are strongly correlated in two modalities obtain higher scores, which is then strengthened by the weight of the self modality that depends on relative rank. Hence, corresponding to each weight graph $\mathcal{R}^{(k)}$, we obtain unified graphs $\mu^{(k)}\ \forall k \in \{p,f,i\}$. These generic unified graphs are further subjected to adaptive unification to obtain cancelable templates.

3) Quality Adaptive Unification: Unified graphs $\mu^{(k)}\ \forall k \in \{p,f,i\}$ are weighted in concurrence with image quality of respective modality. This strategy not only improves system interoperability and recognition rates but also prevents presentation attack. It also adapts the system to a dynamic environment by suppressing low performing modality and simultaneously boosting high performing modality.

For this, we employ a Natural Image Quality Evaluation (NIQE) technique proposed in [35]. NIQE is a blind image quality analyzer that uses a priori knowledge of distortion-free images to construct a general understanding of the natural scene statistics (NSS). Deviation from the general properties indicates degraded image quality. Firstly, training biometric data is fed to extract statistical measures in order to fit a Multivariate Gaussian (MVG) Model $\mathcal{F}_{\text{t}}^{(k)}$ for each modality, $k \in \{p,f,i\}$.

$$\mathcal{F}_{\text{t}}^{(k)} = \frac{1}{(2\pi)^{x/2} |\sum_{\text{t}}^{(k)}|^{1/2}} e^{\left(-\frac{1}{2}(a_x - \varphi_{\text{t}}^{(k)})^T (\sum_{\text{t}}^{(k)})^{-1} (a_x - \varphi_{\text{t}}^{(k)})\right)} \tag{14}$$

such that a_x represents NSS features, $\varphi_{\text{t}}^{(k)}$ denotes the mean and $\sum_{\text{t}}^{(k)}$ represents the covariance matrix of the MVG model. Subsequently, for each query image, statistical features are extracted to fit another MVG i.e. $\mathcal{F}_{\text{q}}^{(k)}$ to compute $\varphi_{\text{q}}^{(k)}$ and $\sum_{\text{q}}^{(k)}$. Distance between $\mathcal{F}_{\text{q}}^{(k)}$ and $\mathcal{F}_{\text{t}}^{(k)}$ is computed to determine the quality score ($\lambda^{(k)}$) for k^{th} modality, $\forall k \in \{p,f,i\}$, defined in Eq. 15. Quality score $\lambda^{(k)}$ is used to weigh corresponding unified vector $\mu^{(k)}$ prior to concatenation to ensure that higher weight is assigned to a more reliable modality.

$$\lambda^{(k)} = \frac{1}{\sqrt{(\varphi_{\text{t}}^{(k)} - \varphi_{\text{q}}^{(k)})^T \left(\frac{\sum_{\text{t}}^{(k)} + \sum_{\text{q}}^{(k)}}{2}\right)^{-1} (\varphi_{\text{t}}^{(k)} - \varphi_{\text{q}}^{(k)})}} \tag{15}$$

Unified template $\mathbf{\Gamma}_q$, is obtained by weighing each unified graph $\mu^{(k)}$ with quality score $\lambda^{(k)}$, followed by concatenation:

$$\mathbf{\Gamma}_q = <\lambda^{(p)} \cdot \mu^{(p)}, \lambda^{(f)} \cdot \mu^{(f)}, \lambda^{(i)} \cdot \mu^{(i)}> \tag{16}$$

1556-6013 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2019.2954779, IEEE Transactions on Information Forensics and Security

Algorithm 2 Adaptive Weighted Graph Fusion (AWGF)

Input: $\mathbf{Q}^{(k)} \in \{\mathbf{Q}^{(p)}, \mathbf{Q}^{(f)}, \mathbf{Q}^{(i)}\}$, κ
Output: Unified Template $\mathbf{\Gamma}$
1: **for** $k \in \{p, f, i\}$ **do**
2: construct $\mathbf{S}^{(k)}$ from $\mathbf{Q}^{(k)}$ using Eq. 10 given κ
3: construct $\mathcal{R}^{(k)}$ from $\mathbf{Q}^{(k)}$ using Eq. 11
4: **end for**
5: **for** $k \in \{p, f, i\}$ **do**
6: compute $\mu^{(k)}$ from $\{\mathbf{S}^{(Z)}\}$, $\mathcal{R}^{(k)}$ using Eq. 12, Eq. 13
7: find $\lambda^{(k)}$ using Eq. 14, Eq. 15
8: **end for**
9: find $\mathbf{\Gamma}$ using Eq. 16
10: **return** $\mathbf{\Gamma}$

TABLE I
MULTIMODAL DATABASES USED FOR EXPERIMENTATION.

Database	Fingerprint	Face	Iris
DB-1	SDUMLA-HMT Multimodal Database [36]		
DB-2	MCYT [39]	CAS-PEAL R1 (expression) [40]	IITD PolyU Iris [42]
DB-3	CASIA-Fingerprint V5 [37]	CASIA-FaceV5 [41]	CASIA Iris V3 (Interval) [43]
DB-4	FVC 2006 [38]	CAS-PEAL R1 (accessories) [40]	CASIA Iris V3 (lamp) [43]

Generated unified cancelable biometric templates are not only dimensionally reduced, but also non-invertible. These templates can be easily revoked either by using a new set of key images, or by changing the order of key images. In addition, quality based adaptation to dynamic environment achieves high distinguishability between presentation attack and mere low quality images. Pseudocode for proposed AWGF is depicted as Algorithm 2. Further, the proposed multimodal biometric system is validated both quantitatively and qualitatively as discussed in following section.

V. EXPERIMENTAL RESULTS AND DISCUSSIONS

A. Databases and Experimental Setup

The proposed multibiometric system is validated using benchmark databases. To validate the performance over a real multimodal database, experiments were conducted using the SDUMLA-HMT Multimodal Database [36], which is dubbed as DB-1 in our setup. Further, following databases have also been been used, CASIA-FingerprintV5 [37], FVC 2006 [38], MCYT [39] for fingerprint, CAS-PEAL [40], CASIA-FaceV5 [41] for face, IITD PolyU [42] and CASIA-IrisV3 [43] for iris. As highlighted by Sultana et al. [10] that correlation among different physiological characteristics could not be established, we formulated virtual multibiometric databases for our evaluation without any loss of generality. For this, unique one-to-one mapping between different unimodal databases is maintained for first *N* subjects to generate three multimodal databases: DB-2, DB-3, and DB-4. Table I summarizes the details about the multimodal databases used. For each subject, one image is randomly selected from five samples and the experiment is repeated five times to achieve 5-fold cross-validation. All experiments were carried out on Windows 10, 2.7 GHz $i7$ processor, 16GB RAM with MATLAB R2018a.

B. Evaluation Metrics

Quantitative performance of the proposed multibiometric system is compared with state of the art fusion methods using performance metrics namely False Match Rate (FMR), False Non-Match Rate ($FNMR$), Genuine Accept Rate (GAR), False Accept Rate (FAR), and Equal Error Rate (EER). In addition, Decidability Index (DI) [44] is determined to gauge the separability between the genuine and impostor score distribution. DI is computed as:

$$\text{DI} = |\mu_g - \mu_{im}|/(\sqrt{(\sigma_g^2 + \sigma_{im}^2)/2}\ ) \tag{17}$$

where μ_g and μ_{im} represent the mean and σ_g and σ_{im} represent the standard deviation of the genuine and impostor score distribution respectively. Further, identification performance is determined using Recognition Index (RI) which is estimated from the top m scores for a query subject.

C. Performance Validation

In this subsection, we validate the system performance from different perspectives namely adaptivity, accuracy, and complexity.

1) Adaptivity Analysis: Adaptiveness of the proposed method is evaluated under the dynamic environment by considering its ability to extract distinctive features from different modalities. For this, we plot the point-set distributions of (a) extracted generic features of individual modality and (b) fused feature to analyze their inter-class and intra-variations. As shown in Fig. 2, normalised features for two multimodal subjects q_1 and q_2 are depicted where $q_1^{(a)}$ and $q_1^{(b)}$ represent two samples of subject q_1. Fig. 2(a) depicts highly distinct features for face modality $Q^{(f)}$, while those extracted from fingerprint $Q^{(p)}$, and iris $Q^{(i)}$, generate a feature with low distinction between subjects q_1 and q_2. In fused features Γ, AWGF efficiently extracts strong correlations between features from different modalities and generates a template which is highly distinct as shown in Fig. 2(b). This is achieved through the cross-diffusion process, where strong information from one modality is fused with information of other modalities. In addition, the generated template has no direct correlation with extracted features as shown in Fig. 2, validating the cancelable properties.

To further evaluate the adaptivity, tolerance of AWGF in (1) *adaptive* and (2) *non-adaptive* modes is compared. By adaptive we refer to AWGF which uses image quality to weigh features, whereas non-adaptive mode has $\lambda^{(k)} = 1, \forall k \in \{p, f, i\}$. For this, performance is compared by adding noise to subsets of input biometric data as tabulated in Table II for DB-2. Wherein, $k' \in \{p', f', i'\}$ represents a noisy modality with, Gaussian noise added with mean $\mu = 0$, and variance $\sigma = 0.01$. Average $EER(\%)$ for single noisy input was observed to be 3.37 and 8.09, and when two inputs were noisy was 5.55 and 9.965 for adaptive and non-adaptive modes respectively. Hence, the adaptive mode performs better with noisy inputs than its counterpart. This behavior is attributed to the fact that adaptive mode adds another dimension, i.e. *image quality* to suppress weak information and ensures high magnitude allocation to more discriminatory information.

1556-6013 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2019.2954779, IEEE Transactions on Information Forensics and Security

TABLE II
IMAGE QUALITY ANALYSIS: COMPARISON OF $EER(\%)$ AND DI FOR ADAPTIVE AND NON-ADAPTIVE MODES OF AWGF. GAUSSIAN NOISE WITH $\mu = 0, \sigma = 0.01$ ADDED TO DIFFERENT SUBSETS OF MODALITIES.

AWGF Mode	**EER**(%)								**Decidability Index (DI)**							
	Original	$(p'fi)$	$(pf'i)$	(pfi')	$(p'f'i)$	$(p'fi')$	$(pf'i')$	$(p'f'i')$	Original	$(p'fi)$	$(pf'i)$	(pfi')	$(p'f'i)$	$(p'fi')$	$(pf'i')$	$(p'f'i')$
Non-Adaptive	1.25	7.72	8.00	8.55	10.05	9.88	12.34	15.12	5.00	2.93	2.90	2.84	2.07	1.85	1.95	1.53
Adaptive	1.02	3.37	2.76	4.00	6.11	5.00	8.78	14.58	5.25	4.14	4.97	3.83	3.47	3.37	2.69	1.68

TABLE III
PERFORMANCE EVALUATION AND COMPARISON. AVERAGE DECIDABILITY INDEX (DI), EER (%) AND RECOGNITION INDEX (RI) FOR FUSED TEMPLATES USING STATE-OF-THE-ART FEATURE FUSION TECHNIQUES AT 95% SIGNIFICANCE LEVEL

Performance Metric	**Decidability** (DI)				**EER (%)**				**Recogniton Index** (RI)			
Method/ Database	DB-1	DB-2	DB-3	DB-4	DB-1	DB-2	DB-3	DB-4	DB-1	DB-2	DB-3	DB-4
CCA-FFS I (*f + i*) [45]	2.21 ± 0.2	2.78 ± 0.8	3.01 ± 0.8	3.11 ± 0.3	6.02 ± 1.1	6.09 ± 1.0	6.00 ± 1.1	5.71 ± 0.6	86.31 ± 2.4	76.31 ± 1.6	91.31 ± 2.3	83.58 ± 1.4
CCA-FFS II (*f + i*) [45]	2.79 ± 0.1	2.66 ± 0.5	2.98 ± 0.9	2.96 ± 0.4	6.34 ± 0.9	7.00 ± 1.1	7.50 ± 1.6	6.02 ± 0.9	79.58 ± 1.3	84.04 ± 1.6	77.22 ± 1.3	82.58 ± 2.1
CCA-FFS I (*f + p*) [45]	3.32 ± 0.3	2.60 ± 0.2	2.36 ± 0.8	2.15 ± 0.5	6.04 ± 0.4	6.46 ± 1.0	5.52 ± 1.1	6.49 ± 0.7	77.13 ± 1.6	76.40 ± 1.9	92.84 ± 1.3	88.87 ± 1.1
CCA-FFS II (*f + p*) [45]	1.99 ± 0.6	2.46 ± 0.4	2.10 ± 0.9	2.25 ± 0.8	7.10 ± 1.0	7.02 ± 1.2	7.00 ± 1.3	6.09 ± 1.2	91.13 ± 1.9	84.13 ± 1.5	79.26 ± 1.5	89.88 ± 2.5
RDM (*f + i*) [20]	3.37 ± 0.1	2.31 ± 0.8	2.28 ± 0.3	2.34 ± 0.3	4.55 ± 0.6	4.00 ± 0.3	5.83 ± 1.0	4.79 ± 1.1	85.28 ± 1.2	87.49 ± 1.6	90.40 ± 1.2	78.72 ± 2.2
RDM (*f + p*) [20]	3.91 ± 0.7	2.42 ± 0.4	3.95 ± 0.5	3.19 ± 0.7	3.96 ± 0.4	6.02 ± 0.2	4.21 ± 0.5	4.67 ± 0.7	89.62 ± 1.4	88.84 ± 1.7	92.91 ± 1.8	87.75 ± 1.1
DCT (*f + i*) [46]	2.43 ± 0.2	2.00 ± 0.4	2.22 ± 0.5	2.08 ± 0.1	5.97 ± 1.1	5.68 ± 0.5	6.22 ± 1.3	7.63 ± 0.7	94.66 ± 1.7	89.35 ± 1.3	92.66 ± 0.9	84.47 ± 0.7
DCT (*f + p*) [46]	2.03 ± 0.4	2.05 ± 0.9	2.47 ± 0.4	2.90 ± 0.3	7.76 ± 1.5	5.96 ± 0.6	5.51 ± 1.1	6.17 ± 0.4	77.96 ± 1.9	93.31 ± 1.6	87.09 ± 1.4	83.59 ± 1.8
Ours (*non-adaptive*)	4.68 ± 0.5	5.09 ± 0.5	4.23 ± 0.4	3.97 ± 1.0	3.12 ± 0.3	1.19 ± 0.1	3.05 ± 0.8	3.44 ± 0.4	96.66 ± 2.1	98.09 ± 0.9	95.88 ± 2.0	91.78 ± 1.3
Ours (*adaptive*)	**4.71 ± 0.3**	**5.38 ± 0.5**	**4.72 ± 0.6**	**4.42 ± 1.1**	**2.03 ± 0.9**	**1.00 ± 0.1**	**1.52 ± 0.5**	**2.35 ± 0.7**	**97.98 ± 1.2**	**99.22 ± 0.5**	**96.66 ± 0.8**	**95.57 ± 1.0**

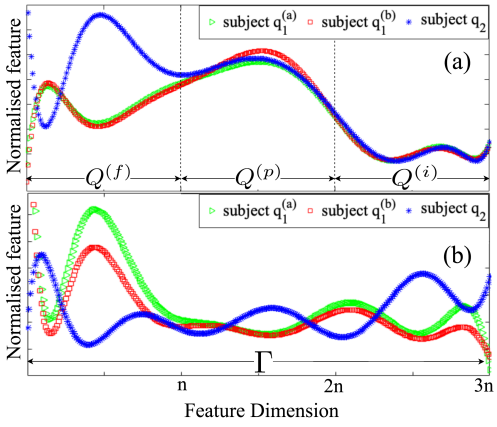


Fig. 2. Feature distinctiveness comparison using point-set distribution of (a) Individual modality generic features, $Q^{(k)}$ and (b) Fused feature, Γ

However, when all three modalities possess noisy information, $EER(\%)$ shoots up to 14.58 and 15.12 for adaptive and non-adaptive respectively. Hence, performance for non-adaptive increase linearly with an increase in noise, while adaptive mode understands the dynamic environments and responds accordingly by differentiating between 'low-quality images' and 'presentation attack'.

2) Fusion Method Comparison: In this subsection, the performance of AWGF in the stolen token scenario is compared with various baseline systems and state of the art fusion methods. Table III reports the accuracy results in terms of average $EER(\%)$, DI and RI at 95% significance level. Further, Detection-Error Tradeoff (DET) curves (Fig. 3) help to analyze the system performance under higher security i.e. at low error rates and support the $EER(\%)$, while CMC curves (Fig. 4) support RI. For comparison, performance of prominent feature fusion methods which can be applied to 1D feature vectors such as Canonical Correlation Analysis (CCA) [45], and Discrete Cosine Transform (DCT) [46] have also been included for different combinations of modalities. Also,

TABLE IV
TIME COMPLEXITY ANALYSIS

Method	**Time (milliseconds)**
RDM	75.0
CCA - I	48.0
CCA - II	49.5
DCT	66.7
Ours (*non-adaptive*)	60.0
Ours (*adaptive*)	61.0

recently proposed transformation based cancelable approach RDM [20] is also evaluated for a comprehensive comparison. To ensure a direct and fair comparison of fusion methods, templates are generated using the same normalized graphs ($Q^{(k)}$) as input features. It is observed that AWGF outperforms the state-of-the-art fusion methods across all databases with an average EER of 1.73%, and RI of 97.35. This is because the proposed method successfully integrates complementary information from different modalities and effectively adapts to information with high reliability. Further, CCA and DCT primarily focus on fusing information without any non-invertibility consideration. Whereas, RDM achieves non-invertibility by using Median Filtering, which is not an inherent property of the fusion process. In contrast, the proposed method generates highly non-invertible templates without any loss in performance.

3) Time and Space Complexity: In this subsection, we analyze the time and space complexity of the proposed approach. Table IV summarises the comparison of time required to generate multimodal biometric templates from the normalized graphs $Q^{(k)}$ as features. Results reveal that the proposed method is computationally efficient due to small feature dimension, which only depends upon the size of the keys. In addition, the proposed method has a significantly smaller space requirement, compared to Gabor features (3.3 MB). Further, the space required by biometric template generated using the proposed method (0.76 KB) is comparable to binary templates for iris like the IrisCodes (1.67 KB) and minutia based fea-

1556-6013 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2019.2954779, IEEE Transactions on Information Forensics and Security

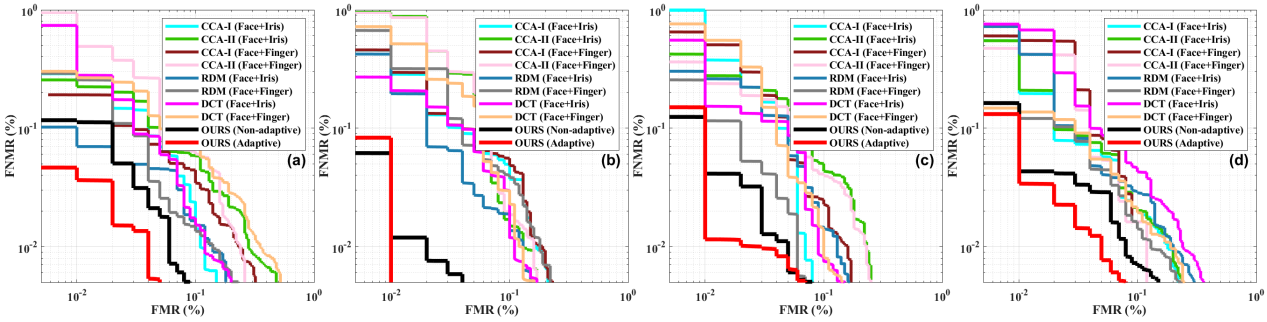


Fig. 3. Detection-Error Tradeoff (DET) Curves, (a) DB-1, (b) DB-2, (c) DB-3, (d) DB-4.

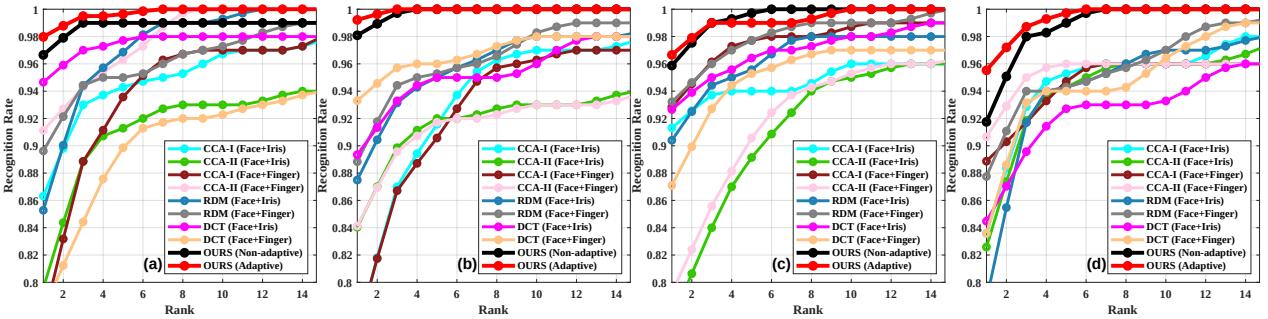


Fig. 4. Cumulative Matching Characteristic (CMC) Curves, (a) DB-1, (b) DB-2, (c) DB-3, (d) DB-4.

tures (0.66 KB). However, comparing minutia based features is not as trivial as comparing similarity-based comparison approaches. In sum, the proposed feature extraction generates generic features with small dimension, which in turn reduces the time complexity for both fusion and comparison process.

VI. Privacy And Security Analysis

Major threats that surround biometric systems involve theft of biometric data, i.e. *privacy* concerns, and illegitimate access, i.e. *security* concerns. The detailed analysis of these concerns are as follows:

A. Privacy Analysis

The privacy concerns of the proposed multimodal biometric system are analyzed through non-invertibility, attacks via record multiplicity, and finally test the unlinkability property of the proposed approach.

1) Non-Invertibility Analysis: Cancelable biometric templates must non-invertible to ensure the privacy of biometric data, in case the biometric system is compromised. In the worst case, the adversary has access to key images. To analyse the non-invertibility, we try to recover the biometric data, i.e. query images $I^{(k)} \in \{I^{(p)}, I^{(f)}, I^{(i)}\}$, from the stored biometric template Γ, and assess the complexity involved. Each element of the template Γ is a real number that may presume any value and has no direct relationship with the key images. Thus, we first assess the complexity of generating λ from template Γ. $(1/\lambda) \in (0, 100)$. To obtain quality scores, one requires the query image itself, making a recursive requirement. Hence, the complexity involved in guessing the quality of all three modalities is $\left(2 \times 10^6\right)^3$, for decimal with 4 places. Further, to invert the fused vector to obtain normalized graphs $Q^{(k)}$, rank graphs $\mathcal{R}^{(k)}$ and sparse graphs $S^{(k)}$ would be required. Assuming rank graphs by brute force, and using Eq. 12 and Eq. 13, relationship between the sparse graphs can be inferred with a complexity of 3n!. To generate sparse graphs, only the parameter κ is unknown, hence to obtain normalized graph $Q^{(k)}$ a complexity of $(n-\kappa)7.5\times10^3$ exists. Further, to generate non-linear graphs $G^{(k)}$, anchors $A^{(k)}$ are required and the complexity involved in generating anchors $A^{(k)}_{avg}$ and $A^{(k)}_{std}$ is $\left(10^4\right)^3$. Given key images, one can correlate them with the non-linear graphs $G^{(k)}$ to obtain biometric data. Table V summarises the worst-case complexity of inverting a template to obtain biometric data i.e. $n!(n-\kappa)1.8\times10^{35}$. Hence, the proposed biometric cancelable templates are highly non-invertible, which cannot be realized in practical scenarios.

TABLE V
Worst Case Complexity Analysis For Invertibility

Generate y from $x : x \rightarrow y$	Complexity	Element Required
$\Gamma \rightarrow \mu^{(k)}$	8×10^{18}	$\lambda^{(k)}$
$\mu^{(k)} \rightarrow \mathbf{Q^{(k)}}$	$3n!$ $(n-\kappa)7.5 \times 10^3$	$\mathcal{R}^{(k)}$ $\mathbf{S}^{(k)}$
$\mathbf{Q^{(k)}} \rightarrow \mathbf{G^{(k)}}$	10^{12}	$A^{(k)}$
Total	$n!(n-\kappa)1.8 \times 10^{35}$	

2) Attacks via Record Multiplicity (ARM): ARM attacks is a more severe attack against the privacy of the biometric data, it uses multiple instances of compromised templates to find a correlation between the templates and biometric data. However, with the proposed biometric fusion, biometric data is cross diffused and transformed into another space where no direct correlation exists between the biometric data and

1556-6013 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2019.2954779, IEEE Transactions on Information Forensics and Security

the templates. Hence, ARM attack complexity is the same as mentioned in Table V, making the system robust against this attack.

3) Unlinkability Analysis: In this section, we evaluate the systems unlinkability using the framework presented by Gomez et. al [47]. Two templates are said to be linkable if the adversary can conclude with certainty that the two templates generated, stem to the same biometric identity. The property of unlinkability is desirable to ensure that the renewed biometric templates and the compromised biometric templates, belonging to the same biometric identity do not correlate. To validate the unlinkability of the proposed system, mated and non-mated pairs are generated. Mated template pairs refer to the biometric templates generated for the same biometric identity using different keys. Whereas, non-mated templates are generated for different biometric identities using different keys. The score distribution obtained by comparing the mated and non-mated template pairs is used for quantitative measurement of unlinkability. The system is said to be unlinkable if the mated score distribution completely overlaps the non-mated score distribution. To evaluate this overlap, we compute the global metric $\mathbf{D}_{\leftrightarrow}^{sys}$, defined in [47]. $\mathbf{D}_{\leftrightarrow}^{sys} \in [0, 1]$, where the system is completely unlinkable if $\mathbf{D}_{\leftrightarrow}^{sys} = 0$, and completely linkable if $\mathbf{D}_{\leftrightarrow}^{sys} = 1$. For this, mated and non-mated template pairs are generated using two scenarios, (a) by using the same set of keys and changing only the order of keys and (b) by changing the key set completely. $\mathbf{D}_{\leftrightarrow}^{sys}$ is calculated for the two scenarios for all four databases DB-1, DB-2, DB-3 and DB-4. Table VI validates the unlinkability of the proposed system wherein unlinkable metric $\mathbf{D}_{\leftrightarrow}^{sys}$ is observed to be close to zero for all four databases. Also, changing the order of keys, and changing the key set completely has the same effect on the unlinkability. This suggests that in order to generate biometric templates, the adversary requires not only the key images but also the order of key images. To further verify that the unlinkability property of the system, we plot the mated and non-mated score distribution in Fig. 5. where the two distributions display a significant overlap. Therefore, we assert that the proposed system supports unlinkability.

B. Security Analysis

Security attacks differ from privacy attacks in the fact that they are meant to gain illegitimate access into the system by either randomly guessing the biometric feature, i.e. (1) brute-force, or using instances of stolen templates to find correlation between template and biometric data, i.e. (2) ARM, or exploit the system's robustness to reject false identities, i.e. (3) false accept attacks. Additionally, (4) substitution attack complexity is assessed to scrutinize the ability of the system to resist

TABLE VI
UNLINKABILITY ANALYSIS FOR TEMPLATES GENERATED BY CHANGING (A) KEY ORDER, AND (B) KEY SET.

Database	$\mathbf{D}_{\leftrightarrow}^{sys}$ (Change Key Order)	$\mathbf{D}_{\leftrightarrow}^{sys}$ (Change Key Set)
DB-1	0.13095	0.12881
DB-2	0.07453	0.09439
DB-3	0.10572	0.09894
DB-4	0.10638	0.11491

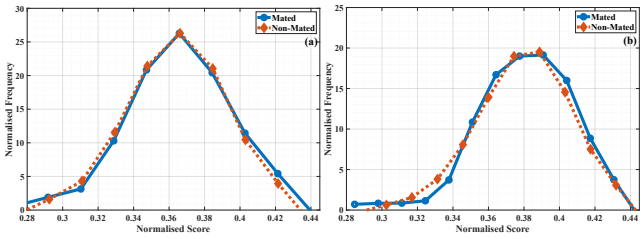


Fig. 5. Unlinkability Analysis: Changing (a) key order and (b) key set

denial of service attacks. Revocability property of the proposed approach is analyzed to ensure that the generated templates are robust. Finally, robustness against presentation attacks is investigated over benchmarked databases.

1) Brute-Force Attack: The attacker in a brute-force attack has no information about the transformation process, keys or the original templates [3]. In a brute-force attack, all possible combinations are tried by the attacker, hoping to guess a legitimate template. For multibiometric data with 3 modalities, template $\Gamma \in \mathbb{R}^{1\times 3n}$ is generated, where n keys are used. Value of each element in the template lies in the range $[0, \infty)$. A real-valued template has infinite guesses and can never be guessed. However, if the range of template values is stolen, an attacker might have a better chance. Therefore, to generate a template we consider the upper limit of the template value equivalent to 1000. It requires $3n^{1000}$ guesses to generate the correct template which is computationally infeasible. To analyze robustness against these attacks, random feature templates were generated and compared with the stored templates for DB-2. Fig. 6 validates the theoretical explanation, where impostor scores were plotted with brute-force attack genuine and impostor scores. Brute-force attack scores overlap completely with the impostor scores and support the claim of the infeasibility of brute-force attacks. Templates generated (a) using image quality and (b) without image quality are both robust to such attacks.

Since, partial knowledge of the template may result in a successful comparison with the enrolled templates, we further analyze the vulnerability of the system against partial matches. For this, we consider the case where an adversary manages to correctly guess $(100-x)\%$ of the template values and rest of the $x\%$ are either unknown or incorrect. Fig. 7, compares the score distribution for genuine, impostor and pseudo-impostor for variable values of $x \in \{20\%, 40\%, 60\%, 80\%, 100\%\}$. Here, $x = 100\%$ represents the ideal brute-force attack case, where no prior information is known about the template. As x increases, the amount of randomly generated (or unknown values) in the template increases. This results in the pseudo-impostor distribution overlapping with the impostor distribution. From Fig. 7, it can be observed that even for low values of x, the score distribution peak for pseudo-impostors is near to the impostor distribution and only a small number of attempts successfully subvert the system by crossing the threshold at the operating point ($FAR = FRR$). Apart from this, even after the assumption of correctly guessing $(100-x)\%$ of the template values, estimation of remaining $x\%$ is quite complex for values of x, as low as 20% is $0.2 \times 3n^{1000}$. Therefore,

1556-6013 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2019.2954779, IEEE Transactions on Information Forensics and Security

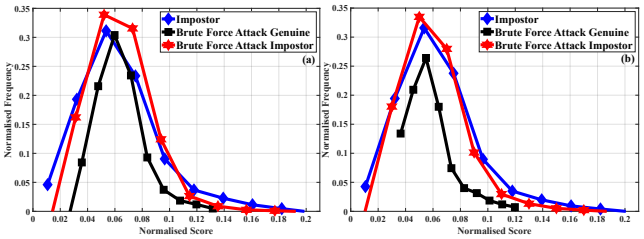


Fig. 6. Brute Force Attack Analysis: Impostor distribution and brute force attack genuine and impostor distribution (a) Adaptive, (b) Non-Adaptive

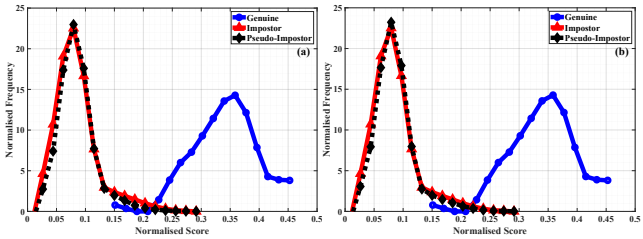


Fig. 8. Revocability Analysis: Plotting impostor, pseudo-impostor and genuine score distributions by changing (a) key order and (b) key set.

the overall complexity for successfully executing a brute-force attack is computationally infeasible due the real numbered template (Γ), generated by a complex diffusion process.

2) Attacks via Record Multiplicity: Attacks via record multiplicity (ARM), also known as correlation attacks [48] utilizes multiple instances of templates that belong to the same biometric identity, but generated using a different set of parameters. Using multiple instances of a template, the attacker tries to determine the correlation between different parameters for reconstructing the original biometric data or the pre-image of the template. Considering two unified feature templates $\mathbf{\Gamma}_1$ and $\mathbf{\Gamma}_2$, which generated using the same biometric data, but different parameters: (a) set of key images, $K_j^{(k)} \in \{1, 2, ..., n\}$, (b) order of key images, (c) κ for KNN, and (d) $\lambda^{(k)} \in \{\lambda^{(p)}, \lambda^{(f)}, \lambda^{(i)}\}$. Unlinkability between these two templates has already been established experimentally. Also, j^{th} value in the template does not correspond to the j^{th} key image and is achieved using a complex computation that involves fusing information from different modalities at varying intensities. Further, another layer of protection is added by weighing the intermediary graphs using the image quality of the query image, and these weights cannot be obtained without the query image. Thereby, making it infeasible to either generate a pre-image or determine the correlation between templates $\mathbf{\Gamma}_1$ and $\mathbf{\Gamma}_2$. Therefore, the complex fusion of information from different modalities along with the involvement of an adaptive weighing technique accounts for the failure of an adversary to utilize multiple instances of templates to find a correlation and forge the system.

3) False Accept Attack: A more sophisticated approach to gain illegitimate access is a false accept attack or dictionary attack, where the attacker is well versed with the template generation process. This increases the odds of generating a legitimate template [48] [9]. To perform a false accept attack, an adversary collects publicly available databases and generates templates using the template generation process.

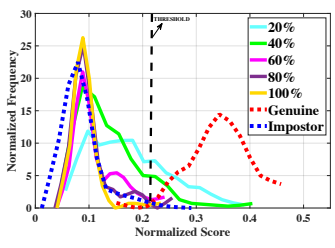


Fig. 7. Score distribution for partial knowledge of template, based on percentage of unknown or incorrectly guessed information about the template.

These pseudo-templates are used to access the system with a probability, equal to the False Accept Rate (FAR). To test the proposed system against false accept attacks, we simulate such an attack using all four databases used for performance evaluation. A subset of one database is generated by randomly selecting m subjects to represent the original template, and the remaining three databases are used to generate pseudo-templates using (a) same set of keys for both original and pseudo-templates and (b) different set of keys for original and pseudo-templates. This process is repeated for each database. Table III. reports the EER(%) for the experiment. For both the key scenarios, the proposed method system displays almost zero FAR, which confirms that the proposed system is robust to false accept attacks. Note that changing the key set does not affect the FAR by much, hence we assert that the proposed fusion process is robust to false accept attacks even for lost key scenarios.

4) Substitution Attack: In such attacks, an adversary may inject its biometric data and replace it with an enrolled biometric record ($\mathbf{\Gamma}_{user}$) [49]. However, the user may or may not possess the knowledge about the algorithms used in a system. Consequently, the enrolled bona fide user may witness a denial of service. This attack is majorly concerned with the security of the database and presents no loopholes in the proposed fusion approach. In AWGF, generating $\mathbf{\Gamma}_{attack}$ from the attacker's biometric data not only requires the knowledge of the complex multibiometric fusion algorithm, but also the complex keys, i.e. key images. In the worst-case scenario where the attacker manages to crack the set of key images along with the order of key images, it would still be a futile exercise. This is because during enrollment all users share the same key images and generate highly distinct biometric templates. Further, generating a combination of all three biometric characteristics (fingerprint, face, and iris) is even more difficult to achieve. Thus, a multibiometric system that fuses information with complex keys is difficult to replicate. Hence, amalgamation with the enrolled user's biometric record would result in a denial of service attack at worst, preventing system

TABLE VII
FALSE ACCEPT ATTACK ANALYSIS

Original Database	Simulated Publicly Available Databases	EER (%)	
		Same Key	Different Key
DB-1	DB-2 + DB-3 + DB-4	0.278	0.264
DB-2	DB-1 + DB-3 + DB-4	0.208	0.215
DB-3	DB-1 + DB-2 + DB-4	0.230	0.205
DB-4	DB-1 + DB-2 + DB-3	0.351	0.273

1556-6013 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2019.2954779, IEEE Transactions on Information Forensics and Security

TABLE VIII
REVOCABILITY ANALYSIS FOR TEMPLATES GENERATED BY CHANGING (A) KEY ORDER, AND (B) KEY SET.

Database	Impostor Mean	Genuine Mean	Change Key Order		Change Key Set	
			Pseudo-Impostor Mean	EER (%)	Pseudo-Impostor Mean	EER (%)
DB-1	0.0869	0.3310	0.0871	2.05	0.0873	2.10
DB-2	0.0851	0.3412	0.0851	1.02	0.0857	1.09
DB-3	0.0852	0.3296	0.0854	1.57	0.0853	1.55
DB-4	0.0879	0.3396	0.0880	2.31	0.0879	2.30

access to both user and attacker. To reinstate the access for the enrolled user, templates can be revoked easily as discussed in the next subsection.

5) Revocability: In this subsection, we analyze the revocability of the proposed approach. Templates are renewed by changing the (a) order of keys, and (b) key set completely. For both scenarios, we plot impostor, pseudo-impostor and genuine score distributions [9] . In this, pseudo-impostors scores are generated by comparing two templates corresponding to the same subject that are generated using different keys. From Fig. 8, it is observed that the impostor and pseudo-impostor distributions for DB-2 show significant overlap, at the same time the genuine distribution is completely offset from the two distributions. This confirms the revocability of the proposed multibiometric system. To further validate the revocability, Table VIII reports the mean of these distributions along with the $EER(\%)$ corresponding to each database. It is interesting to note that changing the order of the keys has the same effect as changing the key set completely. Therefore, the proposed system need not find new key images in case the system is compromised, it just needs to change the order of the key images to revoke the templates.

6) Presentation Attack: Most biometric recognition systems evaluate performance under the ideal scenarios without considering the possibility of presentation attacks. In real-life scenarios, an adversary may supplant the enrolled user's biometric data with an acquired artifact and present it to the biometric capture subsystem in a malicious effort to subvert it.

To achieve reliability against such attacks, while ensuring unhindered access to bona fide users, image quality of the captured biometric data is employed in the adaptive mode of AWGF. In this section, we compare the performance of the non-adaptive and adaptive modes of AWGF to validate the effectiveness of incorporation of image quality against presentation attacks. For this, experiments are conducted using publicly available benchmark databases for presentation attacks, namely, Replay-Attack database [50] for face, Clarkson LivDet-Iris 2017 database for iris [51], and LivDet-fingerprint 2015 database [52] for fingerprint. Three subsets: *train*, *test* and *attack*, were formed such that *train* was used for enrollment, *test* and *attack* were employed as probe biometric data. For evaluation, we compute the Attacks Presentation Classification Error Rate (APCER): ratio of presentation attacks wrongly classified as bona fide or real; and Bona Fide Presentation Classification Error Rate (BPCER): ratio of bona fide presentations wrongly classified as presentation attacks.

Fig. 9 plots the DET curves between APCER and BPCER for both non-adaptive and adaptive AWGF for different combination of spoofed biometric characteristics. The result reveals that for single spoofed modality, the non-adaptive mode observes lower error rates than the adaptive version due to poor quality of spoofed images. In this, adaptive mode reduces the weights ($\lambda_{\text{spoof}}^{(k)} \forall k \in \{p, f, i\}$), and makes the decision based on comparatively high quality captured biometric characteristics. Further, when two modalities are spoofed, the cross fusion of forged information plays a pivotal role, along with the lowered weights for two modalities. Consequently reducing error rates more than the non-adaptive mode. Finally, when all three modalities are spoofed extremely low error rates are achieved, thus ensuring security against presentation attacks. In sum, proposed key-based generic feature extraction in combination with AWGF provides robustness to security attacks and ensures cancelable properties like revocability, unlinkability, and non-invertibility in an effective manner.

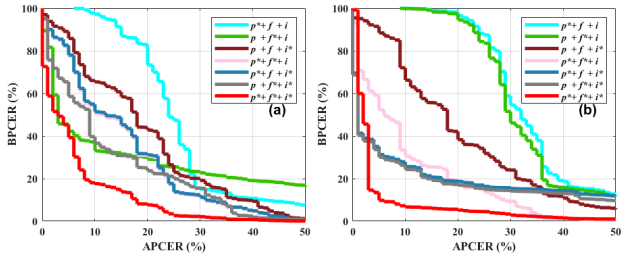


Fig. 9. DET curves for (a) non-adaptive and (b) adaptive AWGF for PAD analysis. k^* = spoofed, k = real, $\forall k \in \{p, f, i\}$

VII. CONCLUSION

In this paper, we have proposed a multibiometric system to achieve high performance along with data protection. Key images based generic feature extraction makes template revocability a facile procedure where simply changing the order of keys results in renewed templates. Also, this drastically reduces feature dimension and hence require low computation and space requirements for realization in real-time. Cross diffusion of rank and sparse graphs extracts complementary information from three modalities, namely fingerprint, face, and iris. AWGF ensures complete non-invertibility and unlinkability of generated biometric templates. Adaptation of features with image quality faithfully resolves the *image quality conundrum*. For this, the non-linear relation between image noise and $EER(\%)$ is exploited to distinguish between 'mere low-quality images' and 'presentation attacks'. The proposed system performs favorably against various security and privacy attacks and hence suitable for security-critical applications.

In future, the proposed fusion approach can be improved to inherently adapt to varying security requirements through a dynamic thresholding mechanism. This would enable the system to operate at an application-specific accuracy-performance tradeoff. The proposed fusion approach displays moderately inferior time complexity compared to other state-of-the-art methods. Hence, this work can be investigated to ameliorate time complexity by carrying out independent computations in parallel. Also, the combination of multibiometric systems and image quality can be investigated for designing a standalone

1556-6013 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2019.2954779, IEEE Transactions on Information Forensics and Security

PAD module. Further, AWGF can be extended to other domains of computer vision bearing multimodal environments.

REFERENCES

[1] A. K. Jain and K. Nandakumar, "Biometric authentication: System security and user privacy." *IEEE Computer*, vol. 45, pp. 87–92, 2012.

[2] N. K. Ratha, J. H. Connell, and R. M. Bolle, "Enhancing security and privacy in biometrics-based authentication systems," *IBM systems Journal*, vol. 40, no. 3, pp. 614–634, 2001.

[3] K. Nandakumar and A. K. Jain, "Biometric template protection: Bridging the performance gap between theory and practice," *IEEE Signal Processing Magazine*, vol. 32, no. 5, pp. 88–100, 2015.

[4] V. M. Patel, N. K. Ratha, and R. Chellappa, "Cancelable biometrics: A review," *IEEE Signal Processing Magazine*, vol. 32, pp. 54–65, 2015.

[5] J. K. Pillai, V. M. Patel, R. Chellappa, and N. K. Ratha, "Secure and robust iris recognition using random projections and sparse representations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, 2011.

[6] M. Gomez-Barrero, C. Rathgeb, G. Li, R. Ramachandra, J. Galbally, and C. Busch, "Multi-biometric template protection based on bloom filters," *Information Fusion*, vol. 42, pp. 37–50, 2018.

[7] A. T. B. Jin, D. N. C. Ling, and A. Goh, "Biohashing: two factor authentication featuring fingerprint data and tokenised random number," *Pattern recognition*, vol. 37, no. 11, pp. 2245–2255, 2004.

[8] Y.-L. Lai, Z. Jin, A. B. J. Teoh, B.-M. Goi, W.-S. Yap, T.-Y. Chai, and C. Rathgeb, "Cancellable iris template generation based on indexing-first-one hashing," *Pattern Recognition*, vol. 64, pp. 105–117, 2017.

[9] Z. Jin, J. Y. Hwang, Y.-L. Lai, S. Kim, and A. B. J. Teoh, "Ranking-based locality sensitive hashing-enabled cancelable biometrics: Index-of-max hashing," *IEEE Trans. Inf. Forensics Security*, vol. 13, 2018.

[10] M. Sultana, P. P. Paul, and M. L. Gavrilova, "Social behavioral information fusion in multimodal biometrics," *IEEE Trans. Syst., Man, Cybern., Syst.*, no. 99, pp. 1–12, 2017.

[11] A. M. Canuto, F. Pintro, and J. C. Xavier-Junior, "Investigating fusion approaches in multi-biometric cancellable recognition," *Expert Systems with applications*, vol. 40, no. 6, pp. 1971–1980, 2013.

[12] G. S. Walia, S. Rishi, R. Asthana, A. Kumar, and A. Gupta, "Secure multimodal biometric system based on diffused graphs and optimal score fusion," *IET Biometrics*, January 2019.

[13] T. Chugh, K. Cao, and A. K. Jain, "Fingerprint spoof buster: Use of minutiae-centered patches," *IEEE Trans. Inf. Forensics Security*, 2018.

[14] J. Galbally, S. Marcel, and J. Fierrez, "Image quality assessment for fake biometric detection: Application to iris, fingerprint, and face recognition," *IEEE Trans. Image Process.*, vol. 23, 2014.

[15] F. Alonso-Fernandez, J. Fierrez, D. Ramos, and J. Gonzalez-Rodriguez, "Quality-based conditional processing in multi-biometrics: application to sensor interoperability," *IEEE Trans. Syst., Man, Cybern. A*, 2010.

[16] N. Poh, J. Kittler, and T. Bourlai, "Quality-based score normalization with device qualitative information for multimodal biometric fusion," *IEEE Trans. Syst., Man, Cybern. A*, 2010.

[17] C. Rathgeb, F. Breitinger, and C. Busch, "Alignment-free cancelable iris biometric templates based on adaptive bloom filters," in *Biometrics (ICB), 2013 International Conference on*. IEEE, 2013, pp. 1–8.

[18] J. Bringer, C. Morel, and C. Rathgeb, "Security analysis of bloom filter-based iris biometric template protection," in *Biometrics (ICB), 2015 International Conference on*. IEEE, 2015, pp. 527–534.

[19] D. Sadhya and S. K. Singh, "Providing robust security measures to bloom filter based biometric template protection schemes," *Computers & Security*, vol. 67, pp. 59–72, 2017.

[20] H. Kaur and P. Khanna, "Random distance method for generating unimodal and multimodal cancelable biometric features," *IEEE Trans. Inf. Forensics Security*, vol. 14, 2019.

[21] H. Kaur and P. Khanna., "Privacy preserving remote multi-server biometric authentication using cancelable biometrics and secret sharing," *Future Generation Computer Systems*, vol. 102, pp. 30 – 41, 2020.

[22] J. Qiu, H. Li, and C. Zhao, "Cancelable palmprint templates based on random measurement and noise data for security and privacy-preserving authentication," *Computers & Security*, vol. 82, pp. 1–14, 2019.

[23] R. Dwivedi, S. Dey, R. Singh, and A. Prasad, "A privacy-preserving cancelable iris template generation scheme using decimal encoding and look-up table mapping," *Computers & Security*, vol. 65, 2017.

[24] D. Sadhya and B. Raman, "Generation of cancelable iris templates via randomized bit sampling," *IEEE Trans. Inf. Forensics Security*, 2019.

[25] J. B. Kho, J. Kim, I.-J. Kim, and A. B. Teoh, "Cancelable fingerprint template design with randomized non-negative least squares," *Pattern Recognition*, vol. 91, pp. 245–260, 2019.

[26] D. Sadhya and S. K. Singh, "Construction of a bayesian decision theory-based secure multimodal fusion framework for soft biometric traits," *IET Biometrics*, vol. 7, no. 3, pp. 251–259, 2017.

[27] W. Yang, S. Wang, J. Hu, G. Zheng, and C. Valli, "A fingerprint and finger-vein based cancelable multi-biometric system," *Pattern Recognition*, vol. 78, pp. 242–251, 2018.

[28] D. Zhong, H. Shao, and X. Du, "A hand-based multi-biometrics via deep hashing network and biometric graph matching," *IEEE Trans. Inf. Forensics Security*, 2019.

[29] I. Chingovska, A. R. Dos Anjos, and S. Marcel, "Biometrics evaluation under spoofing attacks," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 12, pp. 2264–2276, 2014.

[30] S. Wang and J. Hu, "A blind system identification approach to cancelable fingerprint templates," *Pattern Recognition*, vol. 54, pp. 14–22, 2016.

[31] A. S. Chaudhari, G. K. Patnaik, and S. S. Patil, "Implementation of minutiae based fingerprint identification system using crossing number concept." *Informatica Economica*, vol. 18, no. 1, 2014.

[32] M. Haghighat, S. Zonouz, and M. Abdel-Mottaleb, "Cloudid: Trustworthy cloud-based and cross-enterprise biometric identification," *Expert Systems with Applications*, vol. 42, no. 21, pp. 7905 – 7916, 2015.

[33] A. Kahlil and F. Abou-Chadi, "Generation of iris codes using 1d log-gabor filter," in *The 2010 International Conference on Computer Engineering & Systems*. IEEE, 2010, pp. 329–336.

[34] W. Kabir, M. O. Ahmad, and M. Swamy, "Normalization and weighting techniques based on genuine-impostor score fusion in multi-biometric systems," *IEEE Trans. Inf. Forensics Security*, vol. 13, 2018.

[35] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer." *IEEE Signal Process. Lett.*, vol. 20, 2013.

[36] Y. Yin, L. Liu, and X. Sun, "Sdumla-hmt: a multimodal biometric database," in *Chinese Conference on Biometric Recognition*, 2011.

[37] "Casia-fingerprintv5," http://biometrics.idealtest.org/.

[38] R. Cappelli, M. Ferrara, A. Franco, and D. Maltoni, "Fingerprint verification competition 2006," *Biometric Technology Today*, 2007.

[39] J. Ortega-Garcia, J. Fierrez-Aguilar, D. Simon, J. Gonzalez, M. Faundez-Zanuy, V. Espinosa, A. Satue, I. Hernaez, J.-J. Igarza, C. Vivaracho *et al.*, "Mcyt baseline corpus: a bimodal biometric database," *IEE Proceedings-Vision, Image and Signal Processing*, vol. 150, no. 6, pp. 395–401, 2003.

[40] W. Gao, B. Cao, S. Shan, X. Chen, D. Zhou, X. Zhang, and D. Zhao, "The cas-peal large-scale chinese face database and baseline evaluations," *IEEE Trans. Syst., Man, Cybern. A*, vol. 38, pp. 149–161, 2008.

[41] "Casia-facev5," http://biometrics.idealtest.org/.

[42] A. Kumar and A. Passi, "Comparison and combination of iris matchers for reliable personal authentication," *Pattern recognition*, vol. 43, 2010.

[43] "Casia iris image database," http://biometrics.idealtest.org/.

[44] G. O. Williams, "The use of d'as a decidability index," in *Security Technology, 1996. 30th Annual 1996 International Carnahan Conference*.

[45] Q.-S. Sun, S.-G. Zeng, Y. Liu, P.-A. Heng, and D.-S. Xia, "A new method of feature fusion and its application in image recognition," *Pattern Recognition*, vol. 38, no. 12, pp. 2437 – 2448, 2005.

[46] M. I. Ahmad, W. L. Woo, and S. Dlay, "Non-stationary feature fusion of face and palmprint multimodal biometrics," *Neurocomputing*, 2016.

[47] M. Gomez-Barrero, J. Galbally, C. Rathgeb, and C. Busch, "General framework to evaluate unlinkability in biometric template protection systems," *IEEE Trans. Inf. Forensics Security*, vol. 13, 2018.

[48] B. Tams, P. Mihăilescu, and A. Munk, "Security considerations in minutiae-based fuzzy vaults," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 5, pp. 985–998, 2015.

[49] W. J. Scheirer and T. E. Boult, "Cracking fuzzy vaults and biometric encryption," in *2007 Biometrics Symposium*. IEEE, 2007, pp. 1–6.

[50] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," 2012.

[51] D. Yambay, B. Becker, N. Kohli, D. Yadav, A. Czajka, K. W. Bowyer, S. Schuckers, R. Singh, M. Vatsa, A. Noore *et al.*, "Livdet iris 2017iris liveness detection competition 2017," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2017, pp. 733–741.

[52] V. Mura, L. Ghiani, G. L. Marcialis, F. Roli, D. A. Yambay, and S. A. Schuckers, "Livdet 2015 fingerprint liveness detection competition 2015," in *Proc. IEEE 7th Int. Conf. Biometrics Theory, Appl. Syst.*

1556-6013 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

Int J Syst Assur Eng Manag
https://doi.org/10.1007/s13198-019-00926-2

ORIGINAL ARTICLE

An ecological space based hybrid swarm-evolutionary algorithm for software reliability model parameter estimation

Sangeeta[1] · Kapil Sharma[2] · Manju Bala[3]

Received: 15 January 2019 / Revised: 30 October 2019
© The Society for Reliability Engineering, Quality and Operations Management (SREQOM), India and The Division of Operation and Maintenance, Lulea University of Technology, Sweden 2019

Abstract Reliability analysis of the software has attracted a lot of attention of the software developers and researchers due to rapid growing need of software in routine life. The software reliability prediction by mathematical models is entirely centered on the estimation of parameter values, and the parameter estimation of the models poses a non-differential, nonlinear, and multimodal problem. A new algorithm based on the concept of ecological space, method of differential evolution (DE) and intelligent behavior of artificial bee colony (ABC) for optimizing the parameter values has been proposed in this paper. The exploration capability in ABC algorithm has been improved by introducing the concept of ecological space. Ecological space is one of the important factors for evolution and reflects the expansion of individual bee in search space. DE technique provides the diversity of bee's population and faster convergence. The proposed algorithm has been tested with four standard failure datasets. Proficiency of proposed algorithm is also compared with other meta-heuristic algorithms namely ABC, genetic algorithm and particle swarm optimization. Further validation of proposed algorithm is done through comparing its efficiency with hybrid partilce swarm optimization and gravitational search Algorithm. Simulation results verify that proposed hybrid algorithm is very much effective in field of software reliability estimation and would be a competitive one among meta-heuristic optimization algorithms.

Keywords Artificial bee colony · Differential evolution · Ecological space · Genetic algorithm · Particle swarm optimization · Software reliability models

Abbreviations

ABC	Artificial bee colony
DE	Differential evolution
PSO	Particle swarm optimization
GSA	Gravitational search algorithm
GA	Genetic algorithm
SSE	Sum of squared errors
MSE	Mean square error
GO	Goel Okummotto
PTZ	Zhang Tang and Pham model

1 Introduction

Software being the critical component in the growing digital world is playing a key role in every aspect of life whether related to digital identity, digital currency, gadgets, defense, medical, business enterprises, transportation, home security systems or daily money transactions and so on. The software not only improves living standards and

✉ Kapil Sharma
kapil@ieee.org

Sangeeta
sangeeta@dtu.ac.in

Manju Bala
manjugpm@gmail.com

[1] Department of Computer Science, Maharaja Surajmal Institute, Delhi Technological University, Delhi 110042, India

[2] Department of Information Technology, Delhi Technological University, Delhi 110042, India

[3] Department of Computer Science, Indraprastha College for Women, Delhi University, Delhi 110054, India

Published online: 21 December 2019

makes life more comfortable, but also unpredictable software failure creates challenges in life due to reliability concerns of the system. Reliability being the major quality aspect of the software (IEEE 1983) its estimation process must be precise to provide information to the software developers like release time of the software, extent of man-hour consumption etc. Although at most care has been taken while developing a software system using latest tools and techniques but, still it is not feasible to develop defect-free software practically. Software must go through exhaustive testing and debugging, which requires time and money, to enhance the reliability. If the software has faults then it does not mean that it is not reliable, these faults may exist in paths where they are hardly ever executed then that software product is considered reliable. Thus, relationship between fault content in a software, time of fault occurrence and associated software reliability model parameter estimation process must be perceived intensely.

Accurate software reliability estimation is mainly dependent on the selection of optimum parameter values of the models. The techniques for optimizing the parameter values of software reliability models are available through various classical methods of parameter estimation described in Pham (2006) and Myung (2003). These methods are based on the number of constraints and may fall in local maxima and do not converge to global maxima in the multimodal cases. An alternative to these classical mathematical optimization methods is the nature-inspired optimization algorithms proposed by Ab Wahab et al. (2015), Yang (2010a) and Yang (2012b) to solve non-linear, non-differential and multimodal problems.

Nature inspired meta-heuristic algorithms are assembled into four main groups that include algorithms based on evolutionary principles, swarm intelligence behavior, physics phenomena and human intelligence behavior. Holland (1992) and Storn and Price (1997) proposed some algorithms related to the natural evolutionary principle. Chakri et al. (2017), Karaboga and Basturk (2008), Abu-Mouti and El-Hawary (2012), Karaboga and Ozturk (2011) and Eberhart and Kennedy (1995) proposed few swarm intelligence behavior-based algorithms. Kirkpatrick et al. (1983) proposed a physics phenomenon based algorithm. Lim and Isa (2014) proposed an algorithm evolved from the human intelligence behavior. Among these groups, most promising optimization capability is from swarm intelligence and evolutionary principle based algorithms. Darwin's principle-based evolutionary algorithms are like GA, genetic programming and DE. Swarm intelligence behavior based algorithms imitate social activities of creatures like flocks of animals, birds, amphibians and utilizes the social ability of learning and adaptation. These algorithms are the most propitious area of research for numerical optimization and can be successfully applied for parameter optimization of software reliability models. There are numerous such algorithms available, how researchers should use them? To answer this, crucial is to compare these algorithms.

For designing a new algorithm, foremost goal is to know how behavior is pursuing the evolutionary, physics, human and swarm intelligence capabilities. To obtain an optimum solution, there should be an equilibrium between exploration and exploitation geographies of an algorithm. The necessary condition for swarm intelligence is self-organization (comprises feedback, variations and multiple collaborations) and division of labor. For self-organization, fluctuations are vital to provide a level of randomness and for finding the new better solutions effectively. It is important to get rid of stagnation for enhancing the exploitation.

There is a need to find the factors in the algorithm which are playing a key role in calculating the fitness probabilities of the candidate solutions. These factors need to be related to the ecological space fitness as the survival of the fittest is not only the driving factor of evolution as discussed in Sahney et al. (2010).

A new hybrid swarm-evolutionary algorithm for software reliability model parameter estimation is proposed in this paper. The proposed ecological space based algorithm is centered on the social behavior of artificial bee colonies (Yang 2010a; Abu-Mouti and El-Hawary 2012) and evolutionary behavior of DE algorithm (Storn and Price 1997). The swarm intelligence of employee bee is enhanced for providing exploitation to provide a better local search of neighborhood positions using evolutionary principle based DE algorithm. Onlooker bee phase has been improved by incorporating a new factor, showing the fitness probability of the ecological space. Implementation results of the proposed algorithm for software reliability model parameter estimation is showing the validity of proposed algorithm in the field of reliability assessment.

The paper is organized into various sections. Section 1 describes introduction part. Section 2 gives literature review of evolutionary approaches in different domains. Section 3 describes a new proposed hybrid algorithm. Section 4 describes experimental setup. Section 5 discusses results and comparison with other algorithms. The last section provides the conclusion of the paper.

2 Literature Review

For reliability estimation of the software more than 300 software reliability models have been developed by various researchers like (Pham 2006; Goel and Okumoto 1979; Smidts et al. 1998; Zhang et al. 2003; Mahapatra and Roy 2012; Jelinski and Moranda 1972; Musa and Okumoto

1984; Ohba 1984; Yamada and Osaki 1985; Pham and Pham 2019; Majumdar et al. 2019). These models (Roy et al. 2014; Yamada et al. 1986, 1992, 1983; Littlewood and Sofer 1987; Xiang et al. 2015; Yamada et al. 1984; Kapur and Younes 1996; Kumar et al. 2016a, b) are well applicable in a specific type of applications and have used different parameter estimation methods. Parameter estimation process is an important component in estimating the accurate reliability of the software product. Myung (2003), Rana et al. (2013) and Santosh (2015) described some classical methods of parameter estimation for the solution of mathematical optimization problems. Nowadays metaheuristic based optimization methods are found to be performing well in different domains. Algorithms like PSO Eberhart and Kennedy (1995) is inspired by the coordination among bird flocks and GA Holland (1992) is inspired by natural evolution phenomenon, have been well applied in the various fields including the field of software reliability estimation. Haryono et al. (2016) proposed binary PSO algorithm for power system transmission in an optimized way. Jin and Jin (2016) applied improved swarm intelligence approach for parameter optimization. Malhotra and Negi (2013) applied PSO for reliability modeling. Shayeghi et al. (2010) proposed a novel discrete PSO algorithm. Sheta and Al-Salt (2007) applied PSO for parameter estimation of software reliability models. Bi et al. (2015) proposed an improved GA for the optimization of water distribution problem. Ismail et al. (2014) applied GA for the design of hybrid renewable energy systems. Pachauri et al. (2014) applied GA and MAUT in dynamic faults and release optimization based problems. Hsu and Huang (2010) discussed the software reliability models parameter estimation with the applicability of modified GA. Other algorithms like ABC Yang (2012b) have been well applied in various field. ABC algorithm and its variants for constrained optimization have also been discussed in Akay and Karaboga (2017). Karaboga and Ozturk (2011) proposed a new clustering approach using ABC algorithm. Zhu and Kwong (2010) proposed a new global best directed ABC algorithm. Ashish et al. (2018) proposed Parallel bat algorithm-based clustering using mapreduce and Tripathi et al. (2018b) also proposed a new enhanced grey wolf optimizer based clustering method using mapreduce. Karaboga and Basturk (2008) published a research on ABC algorithm's performance. Karaboğa and Baştürk (2007) proposed ABC algorithm to solve problems based on constrained optimization. The food information in ABC is dispersed among the whole bees, this algorithm achieves well in global search optimization i.e. the exploration is upright in finding the better solution. As compared to GA where chances of selection to become a parent depends only on the fitness values of the solution, Storn and Price (1997) proposed DE algorithm in which each solution is having an equivalent chance of being selected as parent, and has been applied in areas like design of IIR filters with requirement of magnitude and group delays, in image segmentation etc. An improved version of DE algorithm generating random number by disordered change has been applied for the safety problem of power system is described in C. and Chen (2006). In this DE algorithm after a few runs may generate overlapping individuals and makes it an inefficient algorithm for global search. Algorithms like Simulated Annealing (Kirkpatrick et al. 1983), Gravitational Search Algorithm (GSA) (Rashedi et al. 2009), Black hole (Hatamlou 2013), Charged System Search (CSS) (Kaveh and Talatahari 2010) and galaxy based search optimization Shah-Hosseini (2011) have been applied in different domains and all these are physics phenomenon based algorithms. Flower Pollination Algorithm (FPA) Yang (2012a) is inspired by the social behavior of pollination in flowers. Cheng et al. (2018) has proposed an Artificial Flora optimization algorithm that can be used for multiobjective, unconstrained multiple variable functions and for combinatorial optimization problems. Bat algorithm Yang (2010b) is based on echolocation feature of bats, have been used for optimization of mathematical problems. A new chaotic multiverse optimization algorithm based on multiverse theory in physics Sayed et al. (2018) has been proposed for solving the engineering optimization problems. Tripathi et al. (2018a) proposed a dynamic frequency based parallel k-bat algorithm for massive data clustering. Chaudhary et al. (2019) proposed a crow search based approach to better estimate the parameters of software reliability models.

Hybrid algorithm have also been applied for the solution of optimization problems like Jadon et al. (2017) proposed hybrid ABC algorithm with DE and applied for optimization of welded beam design problems. This algorithm has provided a modification to position update equation in employee bee phase and they applied DE for onlooker bee phase position update. Mirjalili et al. (2014) proposed a binary optimization for hybrid PSOGSA algorithm. Liu and Zhou (2014) applied improved QPSO (Quantum Particle Swarm Optimization) algorithm in high dimensional complex problems. Li et al. (2013) proposed ABC assisted DE for ORPF (Optimal Reactive Power Flow). They have used ABC in the DE algorithm in-order to recover DE shortcoming of large population requirement for avoiding premature convergence. Abraham et al. (2012) proposed hybrid differential and ABC algorithm. This algorithm applied DE after normal process of ABC for better selection of candidate solution. Mirjalili and Hashim (2010) proposed a hybrid PSO-GSA algorithm by hybridizing PSO and GSA for mathematical function optimization.

Only few algorithms are available in the literature for software reliability model parameter estimation, there is a

need to propose a new algorithm having better exploration and exploitation capability. The capability of the best-featured algorithm in particular domain can be merged for parameter estimation of software reliability models in an accurate and fast manner.

3 Proposed algorithm

A new ecological space based hybrid swarm-evolutionary algorithm has been proposed for software reliability models parameter estimation. ABC algorithm is simple and flexible swarm intelligence based algorithm having fewer numbers of parameters but needs further enhancements to improve its efficiency in terms of exploitation and exploration. ABC algorithm require adaptation for enhancing the exploitation and exploration capability of employee bees and onlooker bees. Food source information shared by the employee bees must be accurate. Employee bees must calculate and share the candidate ecological living space fitness information because ecological space fitness is one of the major factor for evolution of any candidate. There is also a need to incorporate a candidate's ecological living space fitness factor for fitness probability calculations by the onlooker bees so that fitness probabilities could be calculated in precise. This ecological space factor will enhance the exploration capability in ABC.

Proposed hybrid algorithm combines DE algorithm capabilities with ABC algorithm. DE algorithm has the proficiency in improving the diversity of the population by providing an equal probability of being selected as a parent solution from the candidate solutions in space and could better improve local search capability of the employee bee phase in ABC algorithm. DE algorithm also has a faster convergence capability than other evolutionary algorithms and could better fit in ABC algorithm for enhancing its exploitation capability.

Algorithm 1 Ecological space based Hybrid Swarm Evolutionary (DE assisted ABC) Algorithm

```
Initialize S ← swarmsize
Initialize m ← neighborhoodsize
Set the values of Control Parameters
while (Termination Condition) do
  //Employee Bee Phase
  for all particle i in (1,...,S) do
    search the new food source solution for each employee bee in the swarm using Eq. 1

    Go for the neighborhood solution of the current position for each employee bee
    for K0=1-m do
      apply mutation to generate trial vector V^t_{i,k0} as solution using Eq. 2
      Apply Crossover to generate offspring O^t_{k0, d'} using Eq. 3
      do greedy selection for finding the better solution from the current and generated
      solution using Eq. 4
    end for
  end for
  // Onlooker Bee Phase
  for (each of onlooker bee's solution) do
    Based on fitness value probability calculated using Eq. 5 select the food source having
    highest fitness probability.
    Generate neighborhood solution based on the probability calculation and position
    update using Eq. 6.
  end for
  // Scout Bee Phase
  Scout Bee on random, generate the exhausted food source solutions in the space using
  ecological space fitness information and other information from the employee bee in
  the solution space.
end while
Output the Best Solution.
```

The process of hybridization of the proposed algorithm is done in three phases as follows:

1. *Employee Bee phase*
   A level of randomness is essential for maintaining diversity throughout the search space. Quality of the updated solution is influenced greatly by the selection of random solutions. In basic ABC algorithm, the solution is updated only using a random solution of the current search space. Rather than depending only on

this random selection, to maintain population diversity employee bees find updated neighbor-hood positions of the food source from the current positions using DE method, in this each particle breeds locally new and enhanced information from the swarm for finding the neighbor-hood positions. Updated initial position for each employee bee is first calculated using Eq. 1.

$$Y_1{}^{t+1}(i) = P_l(i) + \phi_i(P_l(i) - P_m(i)) \quad (1)$$

And then each employee bee again search for the neighbor-hood solution of the current position for each employee bee using DE. New trial vector is generated, first using mutation and then cross over is done for new offspring generation. At last greedy selection is done from the current and offspring solution for finding better candidate selection. Using Eqs. 2, 3, 4 neighbor-hood solutions are selected.

Generation of m neighbors

For K_0=1-m(neighborhood size) do Mutation

$$V^t{}_{i,k_0} = Y_1{}^t{}_{\mu 1} + C(Y_1{}^t{}_{\mu 2} - Y_1{}^t{}_{\mu 3}) \quad (2)$$

For $d' = 1 - D$ (Search space dimension) do Crossover

$$\text{if } \text{random}_{ij} \leq C_\mu \text{ or } d' = j_{\text{random}}$$
$$\text{Then } O^t{}_{k_0,\,d'} = V^t{}_{k_0,\,d'} \quad (3)$$
$$\text{else } O^t{}_{k_0,\,d'} = Y_1{}^{t+1}{}_{k_0,d'}$$

End for

Selection

$$\text{If fitness } (Y_1{}^{t+1}{}_{k_0}) > \text{fitness } (O^t{}_{k_0})$$
$$\text{Then } O^t{}_{k_0} = Y_1{}^{t+1}{}_{k_0} \quad (4)$$
$$\text{else } O^t{}_{k_0}$$

End for.

2. *Onlooker bee phase*

In the ABC algorithm (Yang 2012b), Onlooker bees selects food sources depending on the information about nectar amount shared by employee bees in the hive. But this is not only the factor for calculating the fitness probabilities of the food source, ecological space fitness information is also one of the driving factor showing the nectar quality and is being shared by employee bees in hive in the proposed algorithm. As per with the state of art algorithms and theories in the literature, ecological living space is a factor that is necessary for the organism (food source) to flourish in the environment and it will enhance the diversity by matching available living space as described in Sahney et al. (2010). The candidate solution and ecological diversity are very much closely related. More favorable is the ecological conditions more quality food will be available to the bees. The diversity will appear to increase depending on the favorable ecological space. The diversity is assessed from ecological, morphological or genetic perspective. Understanding correlation among them can define better the history of the life of candidate solutions on earth.

This is the first application exploring the link between ecological diversity and candidate solution. In the proposed concept candidate food source fitness probability is calculated by onlooker bees using Eqs. (5) and (6).

$$(0.5 * \text{fitness}(i)/\max \text{ fitness}) + \eta_i \quad (5)$$

Further onlooker bees make position update using ecological space factor as given in Eq. (6).

$$Y_1{}^{t+1}(i) = P_l(i) + \phi_j(P_l(i) - P_m(i)) + \eta_i(P_g(i) - P_l(i)) \quad (6)$$

where a factor η is the candidate's available living ecological space fitness probability. Higher fitness probability of the ecological living space will increase the chance of food source as being selected by the onlooker bees.

3. *Scout Bee phase*

Scout bee on random regenerate the exhausted food source solutions in the space using ecological space fitness information and other information from the employee bee in the solution space.

Figure 1 is a two-dimensional space created using the cellular grids. Each position in the cellular grid is a possible candidate solution. Current positions of the food source $Y_{1(i)}$ and updated positions of food source $Y_{1(i+1)}$ are represented by a circle and the circle with plus indicate the possible number of neighbors found using the crossover process. The $Y_{1\mu 1}, Y_{1\mu 2}$ and $Y_{1\mu 3}$ cells in the grid are selected as the possible random candidate solutions and are selected for mutation to generate the V_i vector. This whole process will terminate after a fixed number of iterations and are equal to the swarm size specified.

3.1 Framework of the proposed algorithm

The proposed hybrid Algorithm 1 starts with the initialization of the control parameters, swarm size (S), number of available food sources, limit, the maximum number of iterations, neighbor-hood size (m). Other parameters used in the algorithm are ϕ, C and D. After initialization phase, each employee bee searches for the food source using equation number (1) and then using DE (Eqs. 2, 3 and 4) neighborhood is defined for each employee bee with three

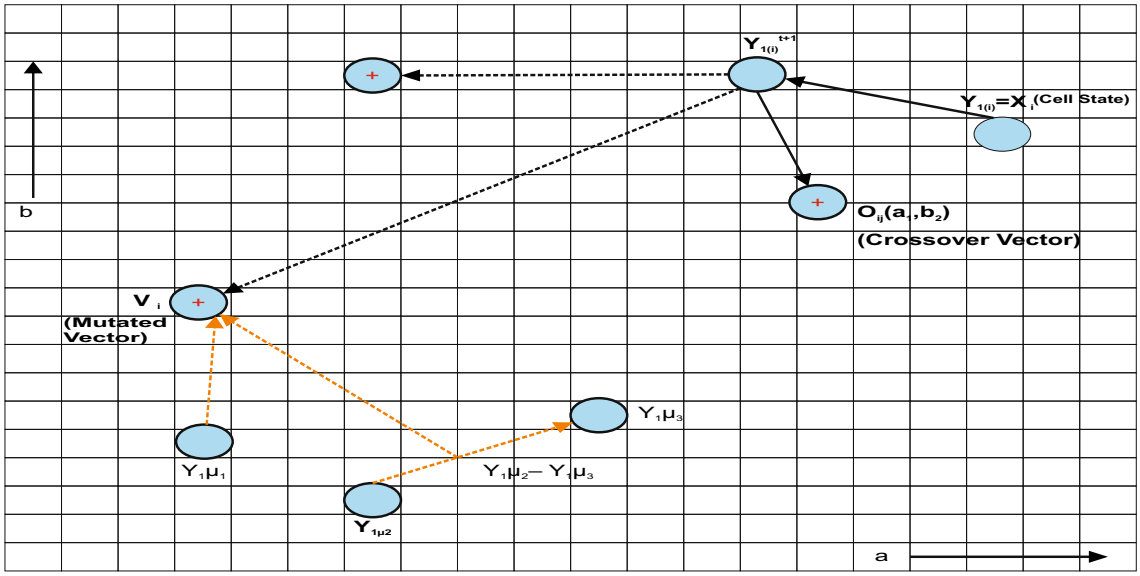


Fig. 1 Ecological space based hybrid DE assisted ABC process

random states in the grid using mutation vector, crossover vector and selection vector.

Then onlooker bee phase takes place that calculates the fitness probability of each state using Eq. (5) and then selects the food source with highest fitness probability. If fitness probability of food source still is not satisfying the termination criterion, then onlooker bee updates positions using Eq. (6). At last scout bee regenerates the exhausted food source in the solution space and finally outputs the best generated solution. Figure 2 depicts diagrammatically flow of the proposed framework.

4 Experimental setup

The implementation of the proposed algorithm is done on Intel(R) Core (TM) i5(5th gen)-62,000 CPU 2.40 GHz with 4 GB RAM and 64-bit windows architecture, x64 based-processor.The proposed algorithm is implemented on most generally used perfect debugging and imperfect debugging behavior based NHPP models. An NHPP is a realistic model for predicting the software reliability. These models considers the debugging process as a counting process characterized by its mean value function.

The performance of the proposed algorithm has been compared with other meta-heuristic algorithms like PSO, GA, ABC and hybrid PSOGSA. Implementation of the algorithms has been done with more than 1000 iterations. Four benchmark datasets shown in Table 1 are used for experimental analysis (Pham 2006). For testing the behavior of proposed algorithm various performance measures are used. These includes estimated model parameter values, Goodness-of-Fit measure of each model in terms of SSE, MSE and convergence behavior in terms of elapsed time in seconds. These measures are calculated and compared with various algorithms using different models and data sets. The best results for each models and data sets are shown in bold for proposed algorithm. The suitability of the algorithm is tested on NHPP models given in Table 2 for experimentation.

5 Results and discussion

5.1 Case 1. Result analysis using GO model and DS1

This case analyzes various meta-heuristic algorithms using two parameter GO model and DS1. Figure 3 is showing the results of expected number of errors i.e. mean value function (mvf) calculated in the software at time t and verifies that the results estimated by hybrid algorithms i.e. PSOGSA and ABCDE are much significant and closer to the actual number of detected errors in the software than other used algorithms. Figure 4 is showing the reliability estimation by various algorithms using GO model on DS1. For proposed algorithm it ranges from 78 to 92% and found to be increasing with time.

5.2 Case 2. Result analysis using Inflection S-shaped model and DS1

Case 2 analyzes various algorithms using three parameter Inflection S-shaped model and DS1. Figure 5 is showing the behavior of various algorithms for mvf analysis and the

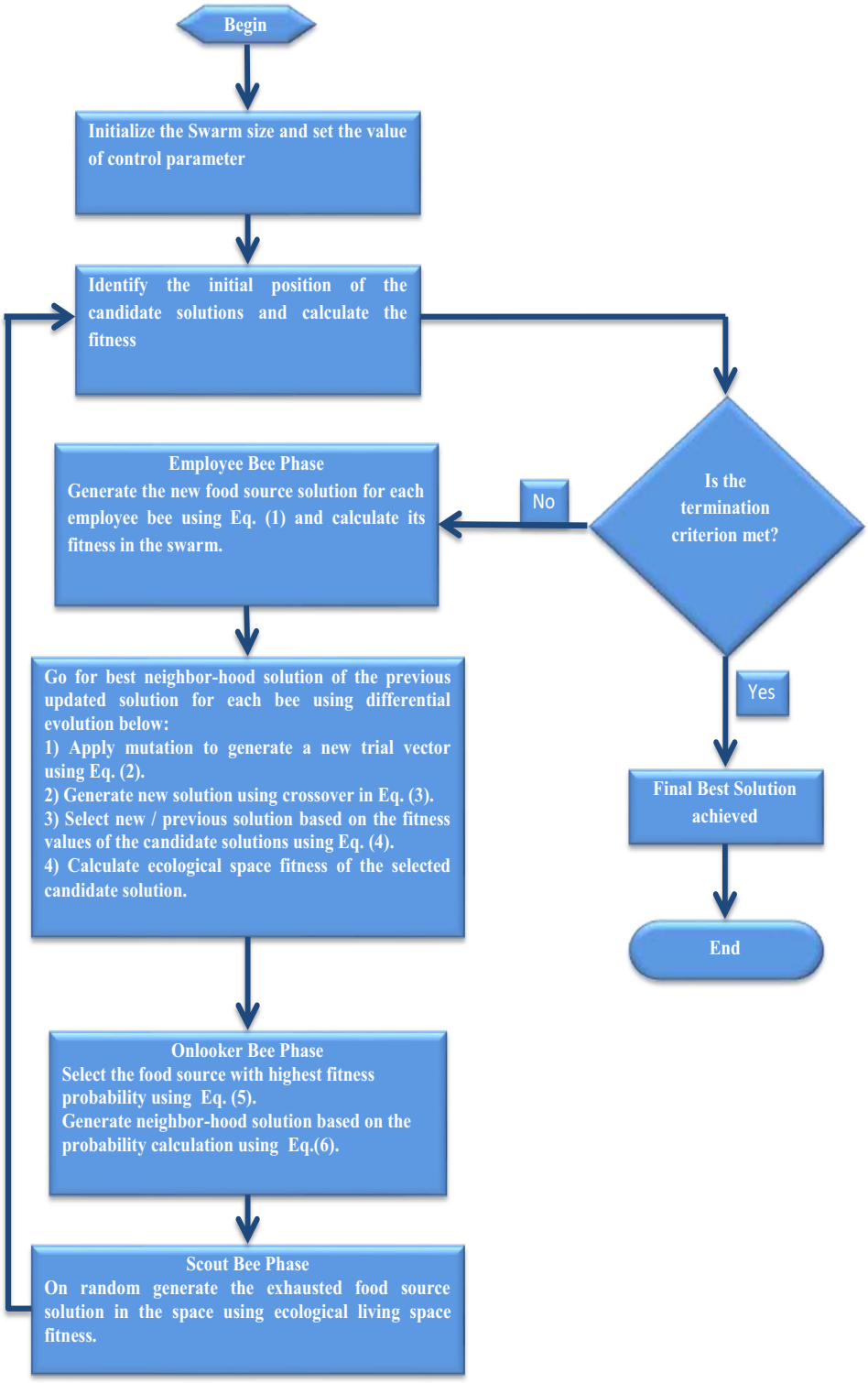


Fig. 2 Flowchart of the proposed ecological space based hybrid algorithm

best results are found using PSO, hybrid PSOGSA and hybrid ABCDE algorithms. Figure 6 is showing the reliability behavior of the software after estimating the errors in the system and is found to be increasing from 79 to 93% in proposed hybrid ABCDE algorithm.

5.3 Statistical measures of DS1

The implementation results for parameter estimation using meta-heuristic approaches for the GO and inflection s-shaped model are given in Tables 3 and 7 respectively and are found to be very much satisfactory. Statistical

Table 1 Datasets used for experimentation

Data set	Number of faults	Time in s/h/days	Type of application	Type of data
DS1	136	25 h	Real time command and control system	Interval domain data
DS2	100	10,000 h	Tandem computers software projects	Interval domain data
DS3	34	849 days	US Navel Tactical data systems	Time domain data
DS4	136	88,682 s	Real time control system	Time domain data

Table 2 Summary of models used for implementation

Model	Behavior	Mean value function
Goel Okumotto NHPP model [17]	Perfect debugging/number of parameter-2	$a(t) = a,\ b(t) = b$ $m(t) = a(1 - e^{-bt})$
Inflection S-shaped model [23]	Imperfect debugging/number of parameter-3	$b(t) = \frac{b}{1+\beta e^{-bt}},\ a(t) = a$ $m(t) = \frac{a(1-e^{-bt})}{1+\beta e^{-bt}}$
Zheng–Teng–Pham model [19]	Imperfect debugging/number of parameter-5	$b(t) = \frac{c}{1+\alpha e^{-bt}}$ $\beta(t) = \beta$ $\mathrm{m(t)} = \frac{a}{p-\beta}[(1 - \frac{(1+\alpha)e^{-bt}}{1+\alpha e^{-bt}})^{\frac{c}{b}(p-\beta)}]$

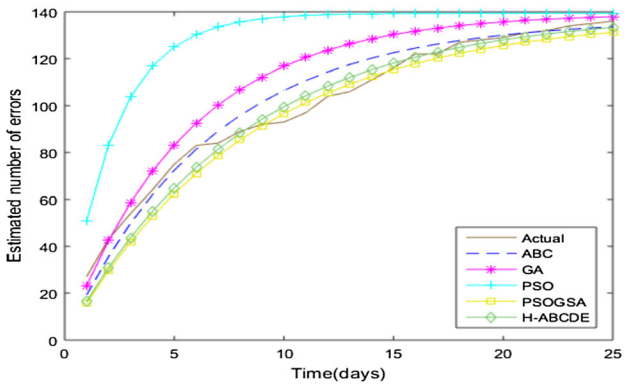


Fig. 3 Expected number of errors using GO model on DS1

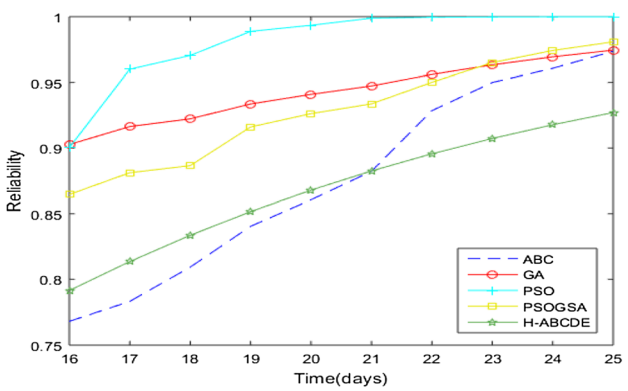


Fig. 4 Reliability estimation using GO model on DS1

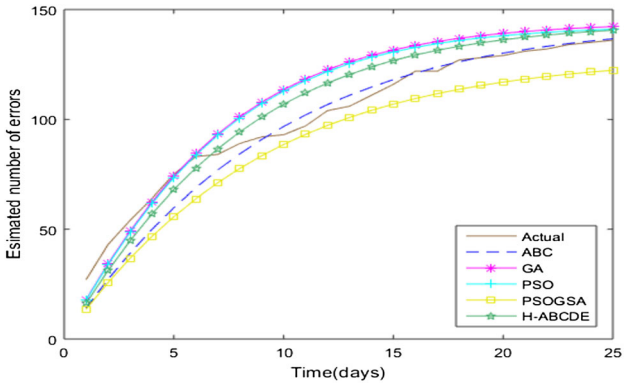


Fig. 5 Expected number of errors using inflection S-shaped model on DS1

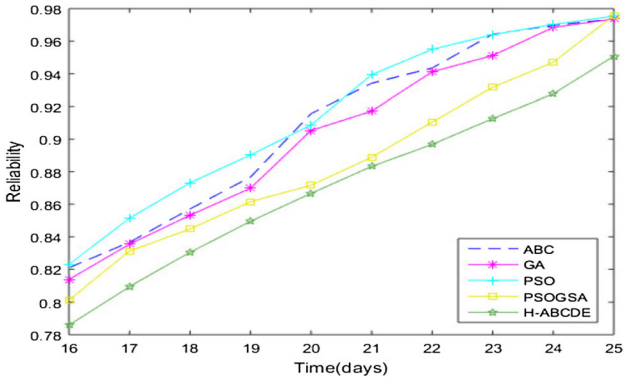


Fig. 6 Reliability Estimation using inflection S-shaped model on DS1

Table 3 Statistical results of parameter estimation for GO model on DS1

Sr. no	Algorithm	a	b
1	ABC	139.4072	0.18173
2	GA	138.9734	0.16789
3	PSO	139.4080	0.18173
4	HPSOGSA	**138.0168**	0.197335
5	HABCDE	**139.4070**	**0.185265**

results of calculated MSE and SSE values are shown in Tables 4, 5, 8 and 9 respectively for both the models. Results are showing that the hybrid algorithms outperforms other algorithms in terms of Goodness-of-Fit measure. Hybrid ABCDE has best MSE and SSE values in average case and satisfactory in other cases as shown in Tables 4 and 5. For inflection s-shaped model best values for MSE and SSE are estimated using the hybrid ABCDE algorithm. Elapsed time showing the convergence behavior of the algorithm is displayed in Tables 6 and 10. The results are showing that the hybrid ABCDE algorithm better converges and do not trap into a local optimum condition, the elapsed time is lesser than GA and little more than others due to a large number of steps taken by the hybrid approach.

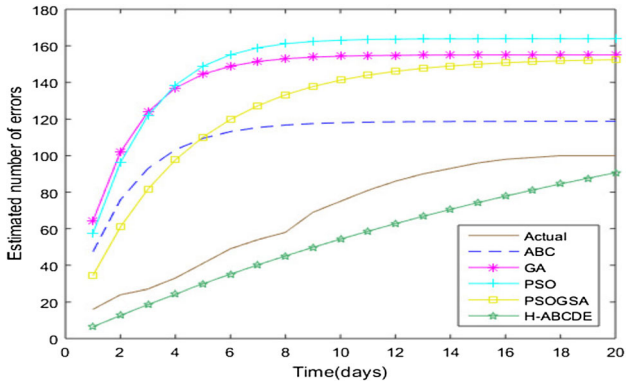


Fig. 7 Expected number of errors using PTZ model on DS2

5.4 Case 3. Result analysis using PTZ model and DS2

In this case analysis of the meta-heuristic nature-inspired algorithms has been done on a more complex model i.e. PTZ model having five parameters on DS2. Estimation of expected number of errors is shown in Fig. 7, hybrid ABCDE algorithm results are very much close to the actual number of errors in the system than other used approaches. Reliability estimated by various algorithms is shown in Fig. 8. It is found to be decreasing at the starting phase and gradually increases when the resources are enough and learning proceeds.

Table 4 Statistical results of mean squared errors (MSE) for GO model on DS1

MSE	ABC	GA	PSO	HPSOGSA	HABCDE
Best (min)	7.16E+02	8.10E+03	1.27E+04	**1.03E+02**	**183.5E+00**
Worst	8.62E+03	1.12E+04	1.53E+04	8.37E+03	**8.05E+02**
Average	2.34E+03	9.99E+03	1.44E+04	4.236E+03	**4.24E+02**
SD	3.51E+03	1092.19	732.8588	3622.776	4.7381

Table 5 Statistical results of sum of squared errors (SSE) for GO model on DS1

SSE	ABC	GA	PSO	HPSOGSA	HABCDE
Best	1790.469	2.03E+05	3.17E+05	**2.58E+02**	4.59E+02
Worst	215,576.4	2.80E+05	3.82E+05	**2.09E+05**	**2.05E+05**
Average	58458.92	2.50E+05	3.59E+05	6.41E+03	**1.26E+03**
SD	87,413.64	27,304.77	18,321.47	90,569.41	99015.88

Table 6 Statistical results of elapsed time in seconds for GO model on DS1

Elapsed time	ABC	GA	PSO	HPSOGSA	HABCDE
Best	27.43	69.6842	4.25219	14.28191	25.45811
Worst	28.00795	84.01981	5.503333	14.64772	30.18599
Average	27.60951	76.87961	4.433742	14.48766	28.25369
SD	0.183413	4.30528	0.346927	0.116742	0.853069

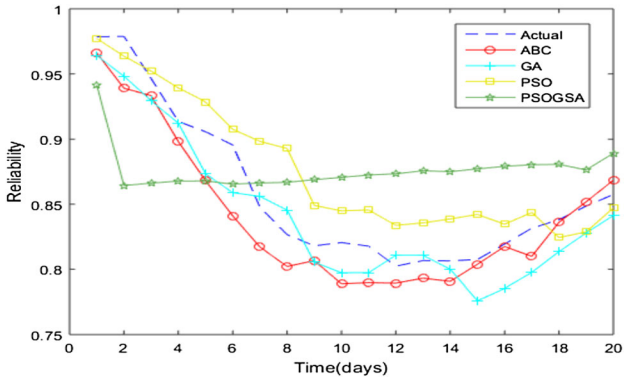


Fig. 8 Reliability estimation using PTZ model on DS2

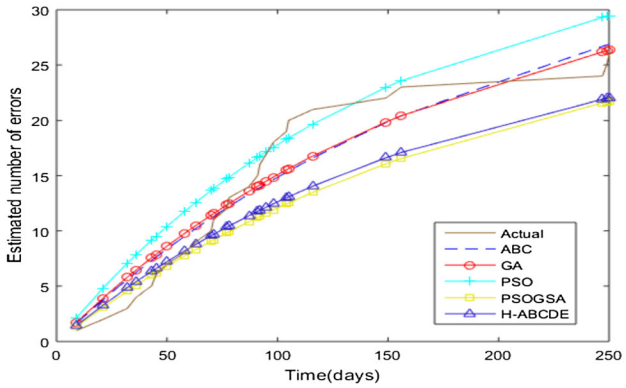


Fig. 9 Expected number of errors using GO model on DS3

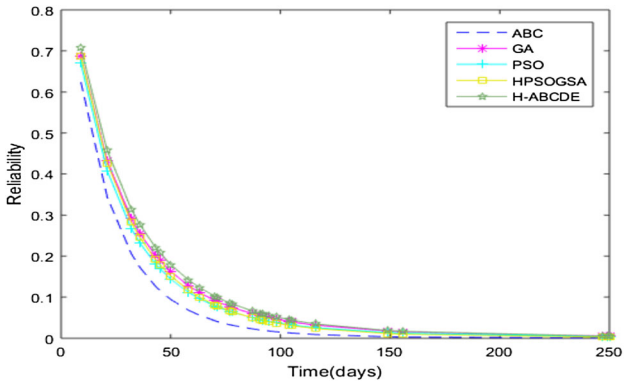


Fig. 10 Reliability estimation using GO model on DS3

5.5 Statistical measures on DS2

Number of parameters estimated by various algorithms is shown in Table 11 and results are found to be satisfactory using hybrid ABCDE algorithm. Tables 12 and 13 are showing significant behavior of hybrid ABCDE algorithm in all three cases using MSE and SSE values. Convergence of the proposed algorithm is shown in Table 14 and found that it does not trap into a local minima or maxima while converging.

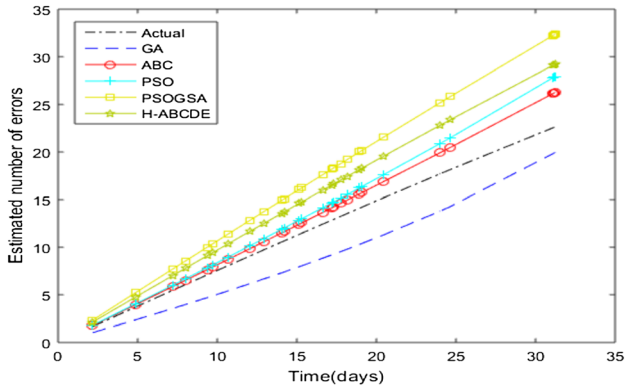


Fig. 11 Expected number of errors using inflection S-shaped model on DS3

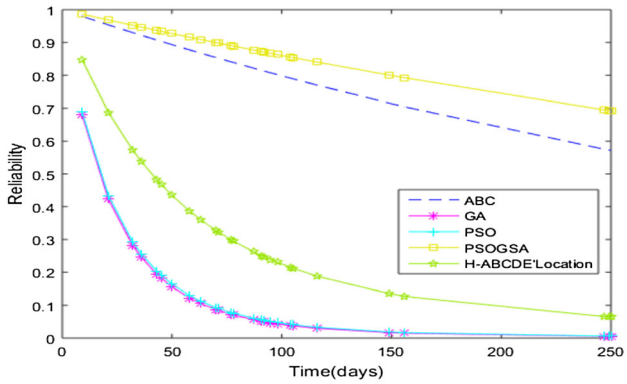


Fig. 12 Reliability estimation using inflection S-shaped on DS3

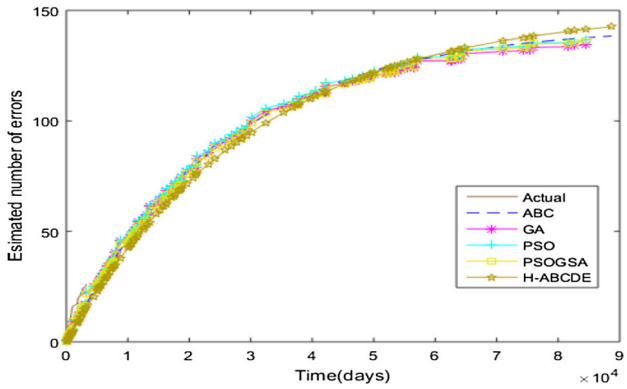


Fig. 13 Expected number of errors using PTZ model on DS4

5.6 Case 4. Result analysis using GO model and DS3

This case analyzes algorithm's behavior using GO model on time domain dataset i.e. DS3. Figure 9 is showing the

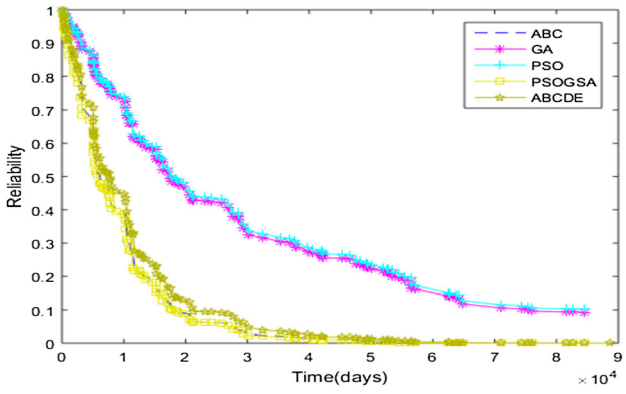


Fig. 14 Reliability estimation using PTZ on DS4

Table 7 Statistical result for Parameter estimation using inflection s shaped model on DS1

Sr. no	Approach used	a	b	Beta
1	ABC	143.5475	0.199477	0.494249
2	GA	139.4072	0.1817297	0.4938127
3	PSO	135.2333	0.197125	0.478044
4	HPSOGSA	**134.0954**	**0.148224**	**0.370924**
5	HABCDE	**138.1621**	**0.159543**	**0.367240**

number of expected errors at time t in the software system using GO model. All algorithms are showing satisfactory results. Figure 10 is showing the reliability of the software system using DS3 and GO model. Highest reliability (up to 72%) is estimated with the hybrid ABCDE method as compared to the other algorithms.

5.7 Case 5. Results analysis using inflection S-shaped model and DS3

This case uses three parameter inflection S-shaped model and DS3 for analysis. A number of errors expected and reliability of the system by various algorithms is shown in Figs. 11 and 12 respectively. Reliability estimated by hybrid PSOGSA is highest among other algorithms. Proposed hybrid ABCDE algorithm estimates upto 85% reliability of the system.

5.8 Statistical measures on DS3

Results for parameter estimation of GO model and Inflection S-shaped model using various algorithms are shown in Tables 15 and 19 respectively. Hybrid ABCDE algorithm performs more significantly among all other algorithms. Tables 16, 17 shows MSE and SSE results for GO model whereas Tables 20, 21 shows MSE and SSE results for Inflection S-shaped model. These results are showing substantial behavior of all the algorithms, but among them for the best and average case behavior of MSE and SSE values hybrid ABCDE algorithm outperforms. The convergence properties are shown in Tables 18 and 22 respectively for both the models. Hybrid ABCDE

Table 8 Statistical result for MSE using inflection s shaped model on DS1

MSE	ABC	GA	PSO	HPSOGSA	HABCDE
Best	8.57E+01	5.17E+03	1.00E+04	4.62E+03	**9.73E+00**
Worst	8.57E+03	1.06E+04	1.06E+04	8.25E+03	**4.56E+03**
Average	2.443E+03	8.941E+03	1.025E+04	6.35E+03	**8.18E+02**
SD	3122.818	2172.934	238.5563	1519.213	1460.607

Table 9 Statistical result for SSE using inflection s shaped model on DS1

SSE	ABC	GA	PSO	HPSOGSA	HABCDE
Best	2.14E+03	1.291E+05	2.450E+05	1.16E+05	**2.43E+02**
Worst	2.14E+05	2.65E+05	2.66E+05	2.06E+05	**1.14E+05**
Average	6.10E+04	2.241E+05	2.56E+05	1.591E+05	**2.04E+04**
SD	78,270.52	54,323.36	5963.906	37,980.32	36,515.16

Table 10 Statistical result for elapsed time in seconds using inflection s shaped model on DS1

Elapsed time	ABC	GA	PSO	HPSOGSA	HABCDE
Best	28.13206	75.68854	4.012182	14.89633	25.25601
Worst	29.00085	78.809	4.290052	15.29576	28.70718
Average	28.55347	77.07568	4.164199	15.01432	26.74783
SD	0.26434	1.460717	0.118187	0.172554	0.401357

Table 11 Statistical result for parameter estimation using PTZ model on DS2

Sr. no	Approach used	a	b	c	Alpha	Beta
1	ABC	108.789	0.0679	0.09664	1.414755	0.000490
2	GA	104.7803	0.090261	1.485479	0.000361	0.000573
3	PSO	109.9867	0.098521	1.692284	0.000389	0.000548
4	HPSOGSA	**104.4378**	**0.069748**	**0.52414**	**0.000386**	**0.00058**
5	HABCDE	**107.7895**	**0.067909**	**0.095645**	**0.0000141**	**0.000491**

Table 12 Statistical result for MSE using PTZ model on DS2

MSE	ABC	GA	PSO	HPSOGSA	HABCDE
Best	5.11E+03	1.97E+04	1.78E+04	**2.241E+03**	**1.59E+03**
Worst	1.84E+04	2.44E+04	2.151E+04	**1.142E+04**	**1.67E+04**
Average	9.36E+03	2.287E+04	1.95E+04	**6.212E+03**	**6.17E+03**
SD	6496.742	2659.868	1187.362	2832.071	5125.092

Table 13 Statistical result for SSE using PTZ model on DS2

SSE	ABC	GA	PSO	HPSOGSA	HABCDE
Best	1.02E+02	3.931E+05	3.57E+05	4.48E+04	3.178E+04
Worst	3.67E+05	4.872E+05	4.29E+05	**2.29E+05**	**3.33E+05**
Average	1.87E+05	4.55E+05	3.90E+05	**1.291E+05**	**1.330E+05**
SD	129,934.8	53,197.36	23,747.25	56,641.42	102,501.8

Table 14 Statistical result for elapsed time in seconds using PTZ model on DS2

Elapsed time	ABC	GA	PSO	HPSOGSA	HABCDE
Best	25.48469	23.29652	2.534641	1.479884	24.00024
Worst	29.67125	26.44875	2.909702	14.8681	26.55558
Average	26.8195	24.37326	2.650514	14.85586	26.2141
SD	1.456462	1.7978	0.055051	42.16803	0.167346

Table 15 Statistical result for parameter estimation using GO model on DS3

Sr. no	Optimization method used	a	b
1	ABC	31.7865	0.0039
2	GA	34.3050	0.0068
3	PSO	35.9554	0.0068
4	Hybrid PSOGSA	**30.0675**	**0.0051**
5	HABCDE	**27.09773**	**0.0066**

algorithms outperforms than ABC and GA algorithm in terms of convergence.

5.9 Case 6. Result analysis using PTZ model and DS4

This case analyzes expected number of errors and reliability estimated by PTZ model and DS4 using different algorithms, these results are shown in Figs. 13 and 14. Hybrid ABCDE algorithm results are very much closer to the actual number of errors in the system than other used approaches. Reliability is found to be decreasing at the starting phase and gradually increases when learning proceeds. From the statistical results in Table 23, calculated for the PTZ model, it is found that hybrid algorithms are

Table 16 Statistical results for MSE using GO model on DS3

MSE	ABC	GA	PSO	HPSOGSA	HABCDE
Best	13.70259	4.582402	4.945849	**4.571759**	**4.64046**
Worst	234.853	19.5016	18.22528	**10.16137**	37.9492
Average	132.7339	13.02272	11.99917	**5.725057**	**11.8947**
SD	79.9948	4.932024	4.459837	1.419409	18.139

Table 17 Statistical result for SSE using GO model on DS3

SSE	ABC	GA	PSO	HPSOGSA	HABCDE
Best	3.556E+02	**1.19E+02**	1.291E+02	**1.193E+02**	**1.21E+02**
Worst	6.11E+03	5.07E+02	4.74E+02	**2.64E+02**	6.19E+02
Average	3.45E+03	3.349E+02	3.12E+02	**1.49E+02**	**2.91E+02**
SD	2079.866	128.2326	115.9558	36.9046	2811.615

Table 18 Statistical result for elapsed time in seconds by various algorithms using GO model on DS3

Elapsed time	ABC	GA	PSO	HPSOGSA	HABCDE
Best	25.15042	50.8989	1.548572	1.696231	26.72567
Worst	38.96231	63.50719	18.13354	2.103597	27.69542
Average	25.82883	53.37236	2.073709	1.877478	26.9957
SD	1.757453	2.7055	2.295692	0.091552	0.247235

Table 19 Statistical result for parameter estimation using inflection S-shaped model on DS3

Sr.no.	Optimization method	a	b	beta
1	ABC	31.77672	7.24E−05	8.80E−05
2	GA	35.77057	0.006895	0.000333
3	PSO	31.08892	0.006895	0.000333
4	HPSOGSA	**31.95481**	**9.54E−05**	**0.000106**
5	HABCDE	**27.28924**	**0.002422**	**7.81E−05**

performing significantly better than other algorithms. The calculated MSE and SSE values in Tables 24 and 25 are showing better performance using GA and hybrid ABCDE algorithms than other used algorithms. Table 26 shows convergence behavior of various algorithms on DS4 in terms of elapsed time.

Expected number of errors in the system and reliability estimation using hybrid ABCDE algorithm are showing good performance of the proposed work. In all cases, proposed algorithm is performing significantly better interms of SSE and MSE values. Calculated elapsed time taken by algorithms for parameter estimation is reflecting the satisfactory convergence behavior of proposed algorithm.

Table 20 Statistical result for MSE using inflection S-shaped model on DS3

MSE	ABC	GA	PSO	HPSOGSA	HABCDE
Best	1.42E+02	**8.25691**	**4.750727**	2.25E+02	15.71997
Worst	2.31E+02	19.46594	**15.11254**	2.33E+02	229.2785
Average	2.08E+02	**14.30301**	**9.784142**	2.30E+02	83.36798
SD	36.0302	4.9158	4.911248	4.137045	78.272

Table 21 Statistical result for SSE using inflection S-shaped model on DS3

SSE	ABC	GA	PSO	HPSOGSA	HABCDE
Best	3.678E+03	**2.15E+02**	**1.24E+02**	5.85E+03	**4.09E+02**
Worst	6.00E+03	5.056E+02	**3.93E+02**	6.056E+03	5.96E+03
Average	5.41E+03	3.72E+02	**2.54E+02**	5.99E+03	**2.15E+02**
SD	936.786	416.4466	127.6924	107.5631	1801.646

Table 22 Statistical result for elapsed time in seconds using inflection S-shaped model on DS3

Elapsed time	ABC	GA	PSO	HPSOGSA	HABCDE
Best	27.58383	79.07232	1.279148	14.52722	27.54145
Worst	40.85063	87.04153	1.502138	15.24965	30.54829
Average	29.92456	81.44662	1.401541	14.72697	28.13046
SD	5.355083	3.2259	0.084879	0.286412	1.13657

Table 23 Statistical result for parameter estimation using PTZ model on DS4

Sr. no	Approach used	a	b	c	Alpha	Beta
1	ABC	139.9816	0.000602	3.99E−05	8.05E−05	9.69E−05
2	GA	138.3598	0.000988	3.98E−05	0.009964	0.000348
3	PSO	139.7796	0.000998	3.99E−05	0.009884	0.007576
4	HPSOGSA	**136.9941**	**0.000933**	**3.30E-05**	**0.008679**	**9.53E−05**
5	HABCDE	**138.4123**	**0.000106**	**3.80E-05**	**0.006291**	**0.00035**

Table 24 Statistical result for MSE using PTZ model on DS4

MSE	ABC	GA	PSO	HPSOGSA	HABCDE
Best	64.29368	**49.98509**	57.25215	58.49756	**56.32027**
Worst	1148.084	**53.31243**	58.89849	94.35258	1558.339
Average	339.4261	**51.4039**	58.11268	77.54565	382.322
SD	407.0099	1.716879	0.700674	15.8634	508.043

Table 25 Statistical result for SSE using PTZ model on DS4

SSE	ABC	GA	PSO	HPSOGSA	HABCDE
Best	8.74E+03	6.80E+03	7.79E+03	7.96E+03	**6.66E+03**
Worst	1.56E+05	7.25E+03	8.01E+03	1.28E+04	**2.12E+03**
Average	4.62E+04	6.99E+03	7.90E+03	1.05E+04	**5.20E+03**
SD	55,353.34	233.4956	95.29173	2157.427	69,093.98

Table 26 Statistical result of elapsed time using various algorithms for PTZ model on DS4

Elapsed time	ABC	GA	PSO	HPSOGSA	HABCDE
Best	45.51209	156.644	51.426	31.22995	50.9900
Worst	46.58381	161.9628	53.83177	31.78577	62.18178
Average	45.91114	159.0327	52.61924	31.472	56.80561
SD	0.406041	2.700436	0.870064	0.206905	1.080029

6 Conclusion

A new algorithm based on the swarm and evolutionary behavior along with the impact of ecological space factor is proposed in this paper. The hybridization has been done in different phases of ABC algorithm. In employee bee phase for finding the better neighbor-hood positions, differential evolution has been used, as DE algorithm outperforms in local search capability. This hybridization will enhance the exploitation capability of original ABC algorithm. In onlooker bee phase for calculating the fitness probabilities of the candidate solutions, a new ecological space factor has been used showing fitness of available living space of the candidate solution. This factor has its importance because survival of fittest is not only the factor for evolution; there is the need of having better living ecological space conditions. This factor has been used further in position update calculation of onlooker bees. Using ecological space factor, exploration capability has been improved as this factor enhances the diversity of the candidate solution. In scout bee phase exhausted food source solutions have been regenerated at random using ecological space fitness information along with other information shared by employee bee in the solution space.

Further comparative analysis of the proposed work with various other nature-inspired algorithms has been done in this paper. The experiment is performed using four real world datasets and well known three SRGMs (GO model, Inflection S-shaped Model and PTZ model). Implementation results (Tables 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26) are showing enhanced capability of the proposed algorithm over the ABC, GA, PSO and hybrid PSOGSA in various means.

Proposed algorithm may have higher complexity in some cases, yet outperforms in other cases. Algorithm shows its convergence in less number of iterations as compared to ABC and GA algorithms.Overall, performance of the proposed algorithm is very competitive.

For future work, proposed algorithm may be used as a generalized algorithm for parameter estimation of other software reliability models. Further proposed algorithm may be applied in other domains for solution of complex engineering optimization problems.

Compliance with ethical standards

Conflict of interest All authors declare that they have no conflict of interest.

Human and animal rights This article does not contain any studies with animals performed by any of the authors. This article does not contain any studies with human participants or animals performed by any of the authors.

References

Ab Wahab MN, Nefti-Meziani S, Atyabi A (2015) A comprehensive review of swarm optimization algorithms. PLoS ONE 10(5):e0122827

Abraham A, Jatoth RK, Rajasekhar A (2012) Hybrid differential artificial bee colony algorithm. J Comput Theor Nanosci 9(2):249–257

Abu-Mouti FS, El-Hawary ME (2012) Overview of artificial bee colony (abc) algorithm and its applications. In: 2012 IEEE International Systems Conference (SysCon). IEEE, pp 1–6

Akay B, Karaboga D (2017) Artificial bee colony algorithm variants on constrained optimization. Int J Optim Control 7(1):98

Ashish T, Kapil S, Manju B (2018) Parallel bat algorithm-based clustering using mapreduce. In: Networking communication and data knowledge engineering. Springer, pp 73–82

Bi W, Dandy GC, Maier HR (2015) Improved genetic algorithm optimization of water distribution system design by incorporating domain knowledge. Environ Model Softw 69:370–381

Chakri A, Khelif R, Benouaret M, Yang XS (2017) New directional bat algorithm for continuous optimization problems. Expert Syst Appl 69:159–175

Chaudhary A, Agarwal AP, Rana A, Kumar V (2019) Crow search optimization based approach for parameter estimation of SRGMs. In: 2019 amity international conference on artificial intelligence (AICAI). IEEE, pp 583–587

Cheng L, Wu Xh, Wang Y (2018) Artificial flora (AF) optimization algorithm. Appl Sci 8(3):329

Eberhart R, Kennedy J (1995) Particle swarm optimization. In: Proceedings of IEEE international conference on neural network. Perth, Australia, pp 1942–1948

Goel AL, Okumoto K (1979) Time-dependent error-detection rate model for software reliability and other performance measures. IEEE Trans Reliab 28(3):206–211

Haryono T, et al. (2016) Novel binary pso algorithm based optimization of transmission expansion planning considering power losses. In: IOP conference series: materials science and engineering, IOP Publishing, vol 128, p 012023

Hatamlou A (2013) Black hole: a new heuristic optimization approach for data clustering. Inf Sci 222:175–184

Holland J (1992) Genetic algorithm. Nat Sci Am 267(1):66–72. https://doi.org/10.1038/scientificamerican0792-66

Hsu CJ, Huang CY (2010) A study on the applicability of modified genetic algorithms for the parameter estimation of software reliability modeling. In: 2010 IEEE 34th annual computer software and applications conference (COMPSAC). IEEE, pp 531–540

IEEE (1983) IEEE standard glossary of software engineering terminology, IEEE STD. 729-19833, IEEE CS Order No. 729

Ismail M, Moghavvemi M, Mahlia T (2014) Genetic algorithm based optimization on modeling and design of hybrid renewable energy systems. Energy Convers Manag 85:120–130

Jadon SS, Tiwari R, Sharma H, Bansal JC (2017) Hybrid artificial bee colony algorithm with differential evolution. Appl Soft Comput 58:11–24

Jelinski Z, Moranda P (1972) Software reliability research. In: Statistical computer performance evaluation. Elsevier, pp 465–484

Jin C, Jin SW (2016) Parameter optimization of software reliability growth model with s-shaped testing-effort function using improved swarm intelligent optimization. Appl Soft Comput 40:283–291

Kapur P, Younes S (1996) Modelling an imperfect debugging phenomenon in software reliability. Microelectron Reliab 36(5):645–650

Karaboğa D, Baştürk B (2007) Artificial bee colony (ABC) optimization algorithm for solving constrained optimization problems. LNCS Adv Soft Comput Found Fuzzy Logic Soft Comput 4529:789–798

Karaboga D, Basturk B (2008) On the performance of artificial bee colony (ABC) algorithm. Appl Soft Comput 8(1):687–697

Karaboga D, Ozturk C (2011) A novel clustering approach: artificial bee colony (ABC) algorithm. Appl Soft Comput 11(1):652–657

Kaveh A, Talatahari S (2010) A novel heuristic optimization method: charged system search. Acta Mechanica 213(3–4):267–289

Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. Science 220(4598):671–680

Kumar V, Mathur P, Sahni R, Anand M (2016a) Two-dimensional multi-release software reliability modeling for fault detection and fault correction processes. Int J Reliab Qual Saf Eng 23(03):1640002

Kumar V, Sahni R, Shrivastava A (2016b) Two-dimensional multi-release software modelling with testing effort, time and two types of imperfect debugging. Int J Reliab Saf 10(4):368–388

Li C, Chen Y (2006) Application of improved differential evolution calculation method based on contious power flow to analyasis of marginal static voltage stability. Power Eng 26:756–760

Li Y, Wang Y, Li B (2013) A hybrid artificial bee colony assisted differential evolution algorithm for optimal reactive power flow. Int J Electr Power Energy Syst 52:25–33

Lim WH, Isa NAM (2014) Teaching and peer-learning particle swarm optimization. Appl Soft Comput 18:39–58

Littlewood B, Sofer A (1987) A bayesian modification to the Jelinski–Moranda software reliability growth model. Softw Eng J 2(2):30–41

Liu F, Zhou Z (2014) An improved QPSO algorithm and its application in the high-dimensional complex problems. Chemom Intell Lab Syst 132:82–90

Mahapatra G, Roy P (2012) Modified Jelinski–Moranda software reliability model with imperfect debugging phenomenon. Int J Comput Appl 48(18):38–46

Majumdar R, Kapur P, Khatri SK, Shrivastava A (2019) Effort-based software release and testing stop time decisions. Int J Reliab Saf 13(3):179–193

Malhotra R, Negi A (2013) Reliability modeling using particle swarm optimization. Int J Syst Assur Eng Manag 4(3):275–283

Mirjalili S, Wang GG, Coelho LdS (2014) Binary optimization using hybrid particle swarm optimization and gravitational search algorithm. Neural Comput Appl 25(6):1423–1435

Mirjalili S, Hashim SZM (2010) A new hybrid PSOGSA algorithm for function optimization. In: 2010 international conference on computer and information application (ICCIA). IEEE, pp 374–377

Musa JD, Okumoto K (1984) A logarithmic Poisson execution time model for software reliability measurement. In: Proceedings of the 7th international conference on software engineering. IEEE Press, pp 230–238

Myung IJ (2003) Tutorial on maximum likelihood estimation. J Math Psychol 47(1):90–100

Ohba M (1984) Inflection s-shaped software reliability growth model. In: Stochastic models in reliability theory. Springer, pp 144–162

Pachauri B, Kumar A, Dhar J (2014) Software reliability growth modeling with dynamic faults and release time optimization using GA and MAUT. Appl Math Comput 242:500–509

Pham H (2006) Springer handbook of engineering statistics. Springer, Berlin

Pham T, Pham H (2019) A generalized software reliability model with stochastic fault-detection rate. Ann Oper Res 277(1):83–93

Rana R, Staron M, Berger C, Hansson J, Nilsson M, Torner F (2013) Comparing between maximum likelihood estimator and non-linear regression estimation procedures for NHPP software reliability growth modelling. In: 2013 joint conference of the 23rd international workshop on software measurement and the 2013 eighth international conference on software process and product measurement (IWSM-MENSURA). IEEE, pp 213–218

Rashedi E, Nezamabadi-Pour H, Saryazdi S (2009) GSA: a gravitational search algorithm. Inf Sci 179(13):2232–2248

Roy P, Mahapatra G, Dey K (2014) An NHPP software reliability growth model with imperfect debugging and error generation. Int J Reliab Qual Saf Eng 21(02):1450008

Sahney S, Benton MJ, Ferry PA (2010) Links between global taxonomic diversity, ecological diversity and the expansion of vertebrates on land. Biol Lett 6(4):544–547

Santosh KG (2015) Numerical methods for engineer, 3rd edn. Age New International, New Delhi

Sayed GI, Darwish A, Hassanien AE (2018) A new chaotic multi-verse optimization algorithm for solving engineering optimization problems. J Exp Theor Artif Intell 30(2):293–317

Shah-Hosseini H (2011) Principal components analysis by the galaxy-based search algorithm: a novel metaheuristic for continuous optimisation. Int J Comput Sci Eng 6(1–2):132–140

Shayeghi H, Mahdavi M, Bagheri A (2010) Discrete PSO algorithm based optimization of transmission lines loading in TNEP problem. Energy Convers Manag 51(1):112–121

Sheta A, Al-Salt J (2007) Parameter estimation of software reliability growth models by particle swarm optimization. Management 7:14

Smidts C, Stutzke M, Stoddard RW (1998) Software reliability modeling: an approach to early reliability prediction. IEEE Trans Reliab 47(3):268–278

Storn R, Price K (1997) Differential evolution-a simple and efficient heuristic for global optimization over continuous spaces. J Glob Optim 11(4):341–359

Tripathi AK, Sharma K, Bala M (2018a) Dynamic frequency based parallel k-bat algorithm for massive data clustering (DFBPKBA). Int J Syst Assur Eng Manag 9(4):866–874

Tripathi AK, Sharma K, Bala M (2018b) A novel clustering method using enhanced grey wolf optimizer and mapreduce. Big Data Res 14:93–100

Xiang J, Machida F, Tadano K, Maeno Y (2015) An imperfect fault coverage model with coverage of irrelevant components. IEEE Trans Reliab 64(1):320–332

Yamada S, Osaki S (1985) Software reliability growth modeling: models and applications. IEEE Trans Softw Eng 12:1431–1437

Yamada S, Ohba M, Osaki S (1983) S-shaped reliability growth modeling for software error detection. IEEE Trans Reliab 32(5):475–484

Yamada S, Ohba M, Osaki S (1984) S-shaped software reliability growth models and their applications. IEEE Trans Reliab 33(4):289–292

Yamada S, Ohtera H, Narihisa H (1986) Software reliability growth models with testing-effort. IEEE Trans Reliab 35(1):19–23

Yamada S, Ohtera H, Ohba M (1992) Testing-domain dependent software reliability models. Comput Math Appl 24(1–2):79–86

Yang XS (2010b) A new metaheuristic bat-inspired algorithm. In: Nature inspired cooperative strategies for optimization (NICSO 2010). Springer, pp 65–74

Yang XS (2012a) Flower pollination algorithm for global optimization. In: International conference on unconventional computing and natural computation. Springer, pp 240–249

Yang XS (2012b) Nature-inspired mateheuristic algorithms: success and new challenges. arXiv preprint arXiv:12116658

Yang XS (2010a) Nature-inspired metaheuristic algorithms. Luniver Press, Bristol

Zhang X, Teng X, Pham H (2003) Considering fault removal efficiency in software reliability assessment. IEEE Trans Syst Man Cybern Part A Syst Hum 33(1):114–120

Zhu G, Kwong S (2010) Gbest-guided artificial bee colony algorithm for numerical function optimization. Appl Math Comput 217(7):3166–3173

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Multimedia Tools and Applications
https://doi.org/10.1007/s11042-019-08495-z

An efficient copy move forgery detection using deep learning feature extraction and matching algorithm

Ritu Agarwal [1] · Om Prakash Verma [1]

Received: 18 August 2018 / Revised: 25 September 2019 / Accepted: 19 November 2019
Published online: 23 December 2019
© Springer Science+Business Media, LLC, part of Springer Nature 2019

Abstract

The image forgery activities are on the rise because of the development of various image editing tools. Such activities are done by attackers with intentions of defaming people and websites or for gaining monetary advantage, extortion etc. Image forgeries are carried out through various ways, among one is the copy-move forgery. The basic process of copy-move image forgery is copying the objects present in an image and create the new image by using the copied objects or placing the copied object on the same image on a different location, hence the need for a forgery detection system to protect the authenticity of images. The existing forgery detection techniques detect the tampered regions with less efficiency because of the large size and lower contrast of the images. This article proposes an efficient technique for detecting the copy-move forged image based on deep learning. The proposed algorithm initializes the tampered image as the input for our system to detect the tampered region. Our system includes processes like segmentation, feature extraction, dense depth reconstruction, and finally identifying the tampered areas. The proposed deep learning based system can save on computational time and detect the duplicated regions with more accuracy.

Keywords Digital image forgery · Copy-move detection · Tampered image identification

1 Introduction

Images nowadays are quite effortlessly manipulated and edited due to the availability of the vast number of image processing and editing software. Forgery is evaluated as the process of molding, adapting, or imitating images and documents. Copy Move Forgery Detection

✉ Ritu Agarwal
ritu.jeea@gmail.com

Om Prakash Verma
opverma.dce@gmail.com

[1] Delhi Technological University, New Delhi, India

(CMFD) is examined as the task of evaluating the image that has been morphologically affected by the application of the corresponding induced mechanism. The CMFD is designed for images that need to be examined for objects as being copied from one area and transferred to another location of the same image. The tampered location of the image is transformed using some distinct transformation such as blurring, rotation, noise formation, scaling and compression respectively [20]. Obliquely, the process of modifying the image so that no optimal tracing method of tampering can be identified. With the development of multimedia technology and multimedia processing tools, the image forgery has been increasingly easy to perform. Although digital images have their benefits, if left unregulated, they more than possess the ability to ruin lives, relationships and government policies. This problem arises because of the ease with which the images can be forged. In addition to the traditional practice, the modification in the images do not vary in accordance with the ancient techniques that are employed in the scheme [1].

The CMFD has been rapidly used in various real time applications, hence the process may be significantly utilized for the authentication purposes applicable in the security areas. To ensure authentication many detection algorithms have been proposed [8, 23, 27]. Nowadays, the technique provides a satisfactory improvement in the process of authentication and is increasing rapidly in the process. Tampering is termed as the scheme that is utilized to import objects over the corresponding image in order to morph the features of the image that are tempted to the destruction of the actual content on the perspective image [4]. To find out the occurrence of the tampering in the image is quite a vigorous task, so much so, that the process may completely edit the features of the image and thereby it will access the original image in order to create the converted copy of the same [17, 19]. The process may create a rapid increase in the operation of the specified issue by reducing the computational complexity using dimensionality reduction [3].

The rate of the occurrence of the forgery has been raising in an adequate amount such that the process can be categorized in many formations, all these cases are classified into three categories based on the process of creating fake images namely, image retouching, and image splicing, and copy-move attack [15, 18]. Image forgery is basically a modification of the image to conceal some meaningful or useful information. The common manipulations of an image are copy-move and cut-paste forgery. Thus the formation of the scheme should provide an adequate response to the corresponding issue of the forgery formation for the same. Thereby the detection process accurately identifies the formation of the forgery over the image and also verifies and overcomes the issue by consistent changes to the image in the same patch itself [11, 16].

The paper is organized as follows. Section 2 of the paper, describe some of the work that already has been conducted to identify the tampered image. Section 3 describes the proposed method for identifying the copied region on images. Section 4 describes the experiments conducted to show the performance of our method and finally, in Section 5 concludes the proposed work.

2 Related work

Nowadays, authentication has been tempted as a vigorous issue in the stream of image detection in accordance with the huge rise in the invention of the technology named image morphing. The authentication and reliability regarding the images are quite detected in a precise manner as such the practice of modifying the images are increasing rapidly.

Subsequently, to overcome the raising of the aforementioned issue, an efficient approach has been induced as CMFD. Warif et al. [21] provides a detailed description of CMF detection techniques and compare them on the basis of block based and keypoint based techniques. In the research, based on the recent deployment in CMFD, the new scheme had been scripted by the researchers. The paper contains the characteristics and various trends of the CMFD. Authors also planned to improve this method by escaping from various new attacks on their future works.

CMFD is termed as an enhancing strategy of depicting the morphological approach to the images in the scheme of user authentication. Henceforth a tremendous research has already been done to overcome the issue and expand the scheme to a next level. Distinct and efficient algorithms can be implemented with the approach to mold the method into an effective one. Bi et al. [5] induced an efficient practice to customize the existing methodology to a capable one, for that a bidirectional coherency error is used to clear the characteristics by iterating the enhanced sensitive gathering. Hence, in accordingly, when the iteration reaches to a specified threshold then the corresponding iteration is quitted in the same duration which rectifies that the deployed feature has been concluded as stable. Conclusively, the obtained results build an optimal outcome on the basis of the development scheme that had been estimated in the research.

Neural Network Prediction is rectified as a process of finding an adequate solution for the examined characteristics of the system. Subsequently, the end-to-end deep neural network can also be intended in the current CMFD approach to finding an accurate solution for the raising issue of duplication of the image in the authentication process. Based on the contribution, Wu et al. [22] lead to an extension of the existing methodology in CMFD to an extreme functioning of the current method. The approach will explicitly introduce a convolutional neural network into the existing technique to extract the feature of the block from the image and also compute the correlation of the different block by matching the point in the persuade feature of the forgery extraction. The process also implements an efficient practice to activate the blurring, modifying and extracting the images. Extensively, the proposed method will lead to an optimal forgery detection that will optimize an accurate outcome when compared with the existing performance in the same.

Multi-scale feature extraction and adaptive matching are tempted as an effective scheme of extracting the morphed characteristic that is intended in an image. The host image is segmented onto the non-overlapping of the irrelevant outlets of the patches in variant scales. Therefore, the corresponding invariant feature extraction of the respective patches are induced in the corresponding application of the perspective method. Adequate patch matching algorithm is being used for finding out the forged region affected in the modified forged image. Additionally, in associated with the Bi et al. [6] introduced a schematic approach in which a subsequent forgery region extraction can be efficiently detected and thus by adapting the methodology will bring an adequate fallout from the practice.

Image tampering is referred to as a unique scheme in image forgery or stealing the images already copyrighted for any person or company. It makes way for the introduction of the copy-move image forgery attack. Thereby the use of the CMFD will often be needed to add and remove the objects from the corresponding image. The forgery analysis of the image is being detected in the modern development of the same based on the effective extraction capability and various transformations of the CMFD respectively. Experimentally, Yang et al. [25] led to an efficient enhancement in the novel featuring of the effective processing in the approached mechanism. Conclusively, key-points are scripted by an improved SIFT descriptor which is

enhanced for the CMFD scenario. The extensive experimental results are presented to confirm the efficiency.

3 Proposed methodology

The simplest method used for faking the image is the copy-move forgery. It can easily be done through various image editing tools. Our method aims to detect such copied regions present in the image. This method is much more convenient even for attacks already have been done by the attackers on the image such as rotation, blurring, and scaling. The flow of the proposed process is shown in the below figure.

Figure 1 describes the process used in our method. Initially, the proposed method uses a tampered image to provide input to the system. After getting the tampered image, the pre-processing process is initiated. In the pre-processing step, the system will segment the input image into several patches based on the similarity of color value. Simple Linear Iterative Clustering (SLIC) is used for segmenting the input image and the features are extracted from the segmented patches for further process. The deep learning method, VGGNet (Visual Geometry Group-net) is used for extracting the multi-scale features from the segmented patches. By extracting features using VGGNet, can efficiently detect copied region even if it rotated, scaled, or it may be compressed. After the feature extraction process, the depth of every pixel in the patches is reconstructed for comparing the blocks. This process is carried out with the adaptive patch matching method. Every segmented patch is compared with one another. During this comparison process, most of the suspicious tampered regions are found and finally, the matched key points are merged with segmented patches of the host image. It exposes the forged region present in the input images. Once the tampered region is identified, it will display it as the output.

3.1 Image segmentation

The image segmentation process carried out by using Simple Linear Iterative Clustering (SLIC). SLIC is a clustering algorithm and works based on K-means clustering algorithm, it separates the area of the similar pixel from the image and these separated pixels are said to as superpixels. Initially, on SLIC, the image will be transmitted to the CIELAB. It contains two feature vectors namely, color value and position of pixels [12].

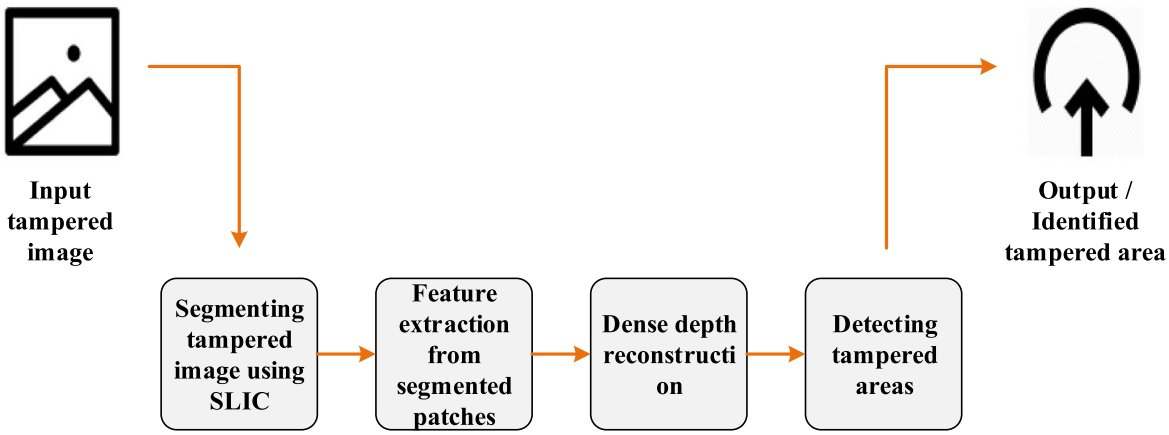


Fig. 1 The process flow of the proposed image forgery detection technique

The color values of pixels in SLIC are given by,

$$Cv = (l, a, b) \tag{1}$$

Where l, a, and b indicates the RGB color values.

The position of pixels in SLIC is given by,

$$P = (x, y) \tag{2}$$

Here, in Eq. (2), x and y indicate the coordinates of the image. The similarity among these 5-D feature vectors will be estimated and the superpixels are segmented by applying local clustering on image pixels.

The size of the superpixel is obtained by dividing the total number of pixels by separated superpixel. Size of superpixel is given by,

$$SSp = \sqrt{n/Sp} \tag{3}$$

Where n indicates the number of pixels and Sp indicates the separated superpixels.

During the clustering process, the seed points in the edge of the image need to be moved to 3×3 region centered by itself, otherwise, it causes some interference in the clustering process.

The similarities of pixels are estimated by using the own seed point and adjacent seed point. The color difference among pixels is estimated by,

$$Dif_{lab} = \sqrt{\left(l_{Sp} - l_i\right)^2 + \left(a_{Sp} - a_i\right)^2 + \left(b_{Sp} - b_i\right)^2} \tag{4}$$

Where, i indicates the corresponding pixels.

The space distance among the pixels is estimated by,

$$Dis_{xy} = \sqrt{\left(x_{Sp} - x_i\right)^2 + \left(y_{Sp} - y_i\right)^2} \tag{5}$$

The similarity between pixels is estimated by,

$$D_i = Dif_{lab} + \frac{m}{S} Dis_{xy} \tag{6}$$

Where, $S = \frac{cv1 - cv2}{255}$, $m = \sqrt{(Spt1 - Spt2)\hat{}2}$

In Eq. (6), where S and $m$describes the relative similarity between color similarity and distance between two seed point (Spt). If the value of D_i is larger, then the two pixels are much related. As the result of this process, the large number of superpixels will be segmented from the input tampered image.

3.2 Feature extraction from segmented superpixels

The superpixels are segmented using SLIC are used for the feature extraction process. The feature extraction from the different segmented patches is carried out by using the VGGNet. The input image [13] that is initialized into the VGGNet need to be in the fixed size of224×224. Our method extracts in six different sizes like 224, 112, 56, 28, 14, and 7. Our feature extraction method includes combination of Convolution and Max-pooling layer. Here, the segmented superpixels are initially binary masked on the input image. A receptive field in image domain contributes every activation of the convolutional feature map so every

activation is projected on the image domain as the center of its receptive field. Every pixels of binary masks on the image is allocated to its adjacent receptive field's center. According to this center and activation location these pixels are projected back onto the convolutional feature map domain. Every location on the feature map will collect numerous pixels projected from a binary mask. Then, the collected binary values are averaged. This will provide a mask on the feature map. This binary mask is multiplied with every channel of feature map to obtain a superpixel features.

Some of the existing forgery detection systems used Scale-invariant feature transform (SIFT), Speeded up robust features (SURF), and some other block based feature for detecting the forged image. The SIFT and SURF features are not efficient while extracting features on low contrast region. The block based methods get low performance while detecting the forgery region because of the higher image size [9]. The proposed VGGNet based feature extraction process can overcome such issues by converting the size of the image when it is initialized into the system and with the increased filters, it extracts more valuable features from the low contrast image.

In Fig. 2, the various layers of VGGNet and its size are described. The input image is initialized and then converted into the size of 224 × 224. From the utilized VGGNet structure, all the 16 layers are combined into several blocks like convolution with max-pooling and the fully connected layer is not shown in this figure. The convolution with max-pooling blocks are composed of 2 or 3 convolution and one pooling layer. The convolution layers in every blocks is responsible for extracting the multi-scale features and the pooling layers is responsible for reducing the size. As a result, it extracts several multi-scale features and the features at max-pooling layers are taken as dense features. The resolution of convolution 1 max-pooling layer is not similar to the input image. Thus, we also extract the features from one layer before the convolution 1 max-pooling layer which has the resolution as same as input image.

In the basic concept of VGGNet, the size of the image is decreased and the number of filtering process are increased during the process. By increasing the filtering process, the best features available in the tampered image can easily be extracted. The extracted features are

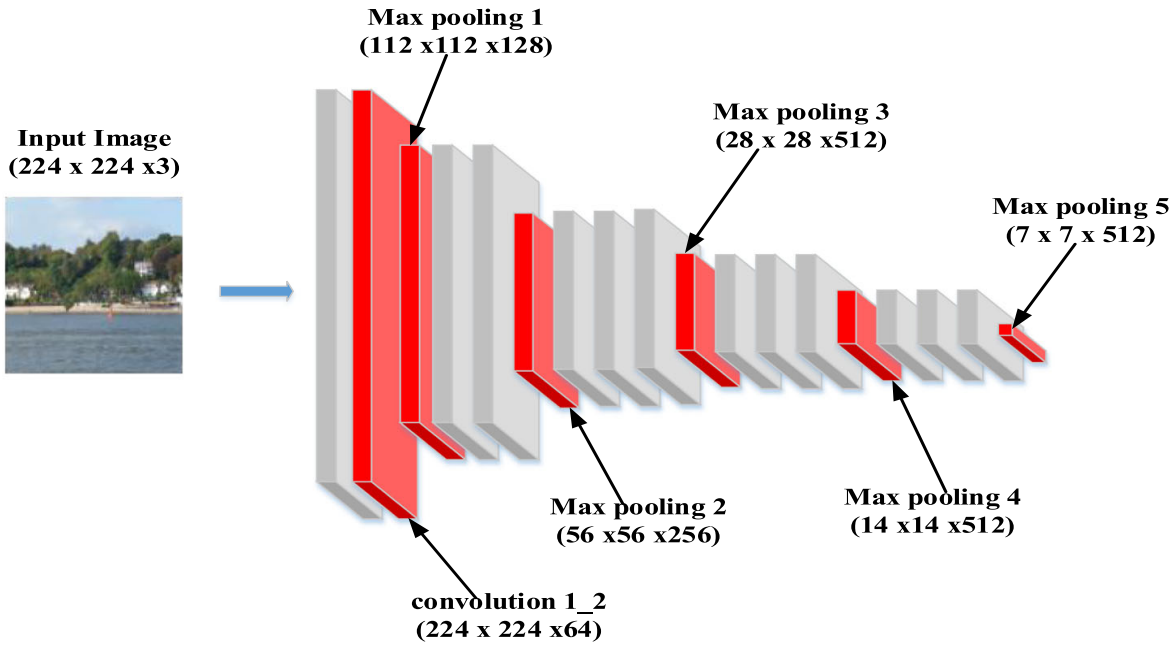


Fig. 2 Features Extracted using VGGNet. The features are extracted from the red colored layers

used for detecting the tampered area in the input image. With the use of convolution and max-pooling the usage of memory and time consuming is reduced in feature extraction process.

3.2.1 Training of VGG for extracting multi-scale features

During the training, the input image is given to the VGGNet via the input layer. The training structure of VGG is characterized by various parameters (more than 340,000) by learning the patches through stochastic gradient descent. The momentum is set to 0.8, the learning rate is set to 0.020 (minimized to 0.010 for every 10 epochs) and the weight decay is initialized to 8.10^{-2}. Here, the first 40 epochs are selected as the trained VGGNet model, because it provides minimal loss on superpixel validation. The training parameters utilized for VGGNet is given in Table 1.

The single scale features from image cannot provide more detailed information and minor variation among the original and tampered region. So, here multi-scale features are extracted with VGGNet, it extract all the regions with similar information (even with minor variation). While sending these features for depth reconstruction, it easily provide most continuous matching pixels present in the image and it also helps to identify the tampered region more accurately.

3.3 Dense depth reconstruction

In this section, the dense depth of the patches is reconstructed on the trained model and then matched with other patches for identifying the suspicious tampered region. By comparing the patches, the tampered areas available in the image are visible and can be identified. To achieve accurate reconstruction with highly repeatable dense VGG features, the keypoint locations are relocalized in order to give efficient pixel-level feature correspondences. The initial correspondence can be established by using upper max-pooling layer. Because, higher receptive field is in the upper layer and more semantic data are encoded by this layer. To preserve pixel-level quantity and accuracy, the following procedure is used for relocating the key points. For every feature at the present layer, the 2×2 neighborhood VGG descriptors are retrieved on the lower layer (larger spatial resolution). This feature point is relocated at the pixel location which consumes the highest descriptor norm (L2 norm) in the 2×2 neighborhood. That is, a map

Table 1 Training structure of VGGNet

Blocks and layers		Size			
		Input	Filter	Kernel	Output
Convolution with max-pooling 1	Convolution layers	224	64	3 × 3	112
	max-pooling layer	–		–	–
Convolution with max-pooling 2	Convolution layers	112	128	3 × 3	56
	max-pooling layer	–		–	–
Convolution with max-pooling 3	Convolution layers	56	256	3 × 3	28
	max-pooling layer	–		–	–
Convolution with max-pooling 4	Convolution layers	28	512	3 × 3	14
	max-pooling layer	–		–	–
Convolution with max-pooling 5	Convolution layers	14	512	3 × 3	7
	max-pooling layer	–		–	–

generated at lower level whose spatial resolution is high is used to relocalize the keypoint on a sparser level. This process is terminated when it reaches the convolutional layer which has the same resolution as the input image. An example for key point relocalization process for the particular feature point at the final feature map of VGGNet is illustrated in Fig. 3. This Figure shows the relocalization processes until the convolution with max-pooling 3 and this process should be repeated for remaining layer. The overview of depth reconstruction is shown in Fig. 4:

It need to create the depth map for the objects in the images based on the relocated features for initializing the reconstruction process. The depth map DM is created based on the features vectors obtained from VGG net and is given by,

$$\mathrm{DM} = \left\{ \mathrm{F}_{\mathrm{i,j}} | \mathrm{i} \in 1 \ldots \mathrm{w},\ \mathrm{j} \in 1 \ldots \mathrm{h} \right\} \tag{7}$$

Where, $\mathrm{F}_{\mathrm{i,\ j}}$ represents the color values or the brightness of pixels extracted as feature at location i and j, and w and h represent the width and height factor of feature vector respectively.

Implicitly the DM is defined as $\Delta \mathrm{DM} = 0$, the Δ represents the Laplace operator. Based on this, the depth for pixels (i, j) are determined and it is given as,

$$4\mathrm{DM}_{\mathrm{i,j}} - \mathrm{DM}_{\mathrm{i+1,j}} - \mathrm{DM}_{\mathrm{i-1,j}} - \mathrm{DM}_{\mathrm{i,j+1}} - \mathrm{DM}_{\mathrm{i,j-1}} = 0 \tag{8}$$

For reconstructing the dense depth of the tampered image, Median difference dif_{xy} is calculated on all patches of created depth map by using the extracted features and compared with the Median threshold *th*. If the difference rate is smaller than the threshold, then the newer difference $Ndif_{xy}$ is calculated. The new difference value is estimated by,

$$Ndif_{xy} = |In_x - In_y| \tag{9}$$

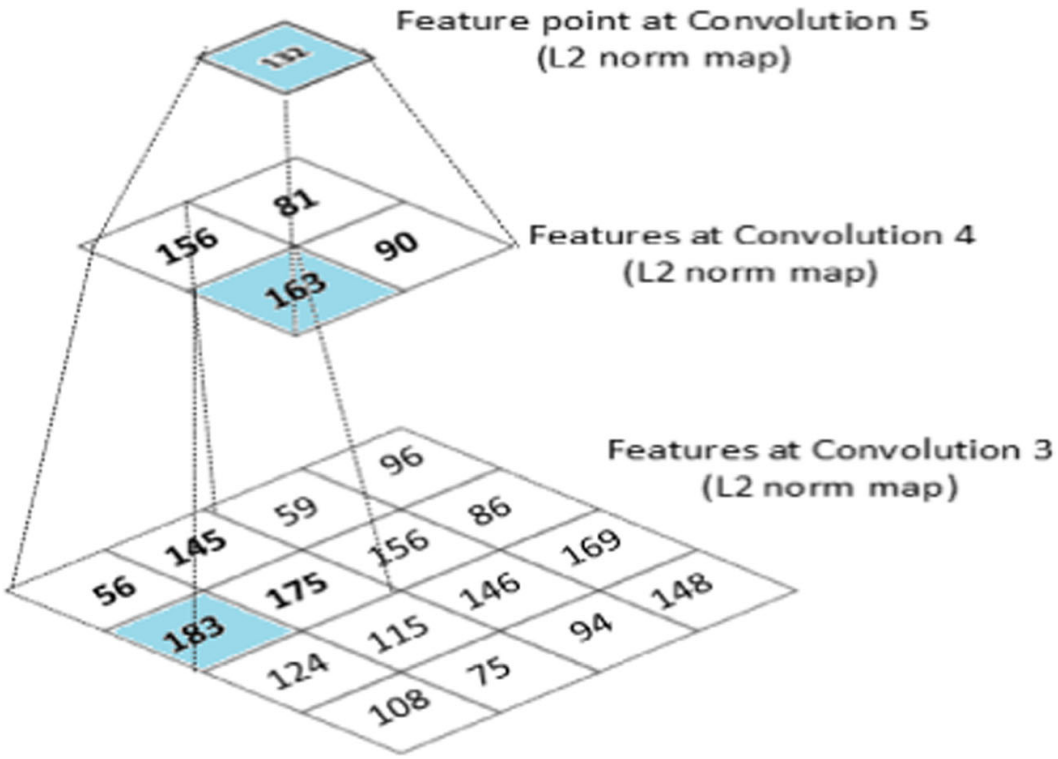


Fig. 3 Example for Keypoint relocalization

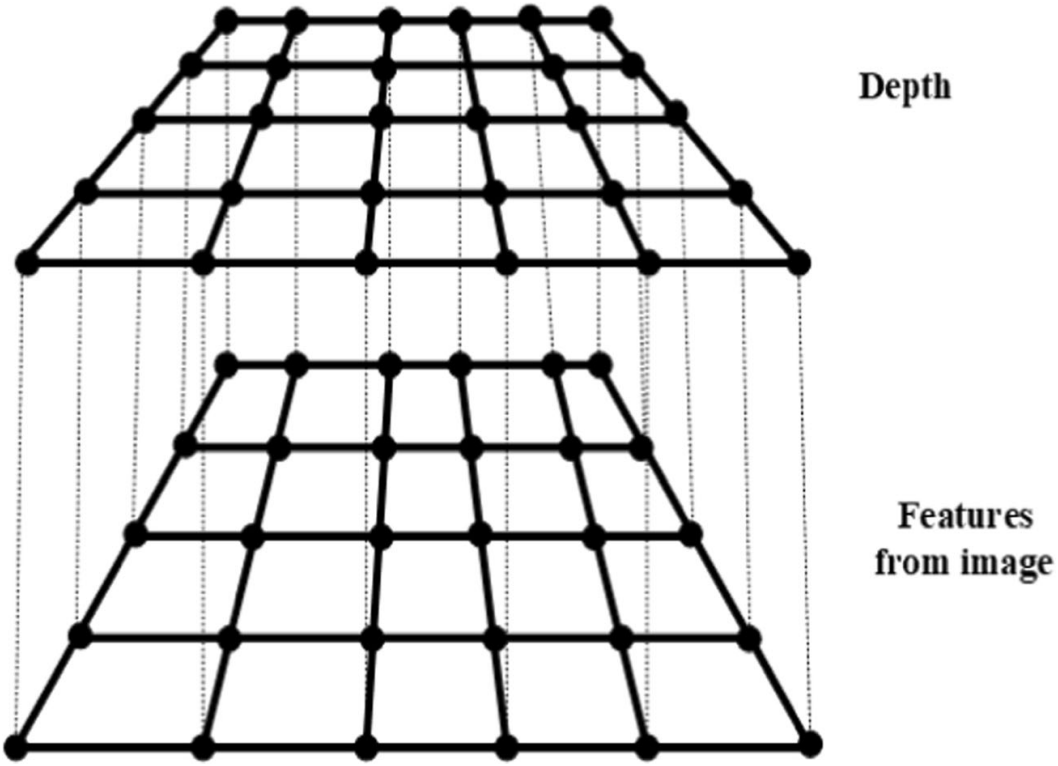


Fig. 4 Overview of depth reconstruction with extracted features

In Eq. (9), In_x and In_y are the intensities of various patches. The similarity among the patches will be visible by reconstructing the patches. For this, every patch is converted into binary and it is given by,

$$Ndif'(i,j) = \begin{Bmatrix} 0 & Ndif_{xy} \leq th \\ 1 & Ndif_{xy} > th \end{Bmatrix} \tag{10}$$

From eq. (10), the zeros detected from the binary block is used to determine the similarity among the patches. At this step, most of the suspicious regions are identified with the connected zeros. The binary conversion of the patches helps to identify the tampered region more accurately and efficiently. The determination and reconstruction of depth value reduce the higher dissimilarity among the original and forged patches.

3.4 Matched region identification with APM

Here, by using the Adaptive patch matching (APM) algorithm [7], the similar key points available in every patch are extracted.

Initialize the reconstructed patches,

$$RP = \left\{RP^1, RP^2, ..., RP^n\right\} \tag{11}$$

In Eq. (11), the number of scales is described as n.

After the initialization, the correlation among the patches is estimated. It is done by matching the key points of patches and it is given by,

$$CRP = \left\{CRP^1, CRP^2, ..., CRP^n\right\} \tag{12}$$

The *CRP* indicates the correlation coefficients of every patch.

Based on the estimated *CRP*, calculate the threshold for every patch. Which is given by,

$$th_{RP} = \left\{ th_{RP}^{1}, th_{RP}^{2}, ..., th_{RP}^{n} \right\} \tag{13}$$

The similar patch pairs are found by using the threshold of every patch and it is described as,

$$SP = \left\{ SP^{1}, SP^{1}, ..., SP^{n} \right\} \tag{14}$$

The matched key points from the patches are extracted using similar patch pairs and it is given as,

$$MKP = \left\{ MKP^{1}, MKP^{2}, ..., MKP^{n} \right\} \tag{15}$$

3.5 Merging matched keypoints and forgery detection

In this section, the keypoints extracted from the previous step is merged with segmented patches to detect the forged region. For this, the host image is segmented into several patches, after that the matched keypoints *MKP* is merged to the segmented patches. By merging the *MKP* with the segmented patches can remove the unforged area and expose the forged area more efficiently.

The time period of pixel appearance in each scale is initially calculated and it is given by,

$$TP = \left\{ tp_{\min}, tp_{\min+1}, ..., tp_{\max} \right\} \tag{16}$$

Where " min " and " max "indicates the minimum and maximum rate of pixel appearance *TP* in all scales. The *TP* always be in random sequence because of the various input image and it also satisfies the condition$tp_{\max} \leq n$. By using the mean M and standard deviationSD, the unforged regions are removed. It is estimated by,

$$M = \frac{1}{\max - \min} \sum_{i=\min}^{\max} tp_i \tag{17}$$

$$SD = \sqrt{\frac{1}{\max - \min} \sum_{i=\min}^{\max} (tp_i - M)^2} \tag{18}$$

By subtracting the mean value with twice the value of Standard deviation, the unforged area can be removed and it is given as merging threshold ($M - 2SD$) (i.e., if the appearance of a pixel time period is lesser than the merging threshold, then that pixel is removed).

The merging process of matched keypoints with the segmented patches is given by,

$$MR(x,y) = \begin{cases} 1 & \mathrm{M-2SD} \leq \sum_{i=1}^{n} f_i(x,y) \leq tp_{\max} \\ 0 & 0 \leq \sum_{i=1}^{n} f_i(x,y) < M - 2SD \end{cases} \tag{19}$$

In eq. (19), the merged region is described as$MR(x, y)$, suspected tampered/forged region in the i^{th}scale is described as$f_i(x, y)$, and n indicates the number of scales. Finally, the key points are merged and the forged regions of the tampered host image are identified.

4 Experiments and discussion

This section explains the performance of the proposed deep learning based method. The experiments are conducted in the MATLAB tool with MICC-F220 dataset. The dataset includes 220 images (110 tampered and 110 forged), which contains various types of attacked images such as, rotated, scaled, and noised.

Performance evaluation of comparison process is done with two existing systems, Invariant feature based technique [2], SWT (Stationary Wavelet Transform) feature based technique [14], hybrid feature based technique [24], and MROGH feature based technique [26]. The paper compares the performance in terms of True Negative Rate (TNR), False Negative Rate (FNR), False Positive Rate (FPR), Precision, Recall, F1-Score, Accuracy, and E-measure.

This part initially describes the details about the performance metrics which is used for comparison (4.1), different attacks that are used in copy-move forgery techniques and its results such as segmented image and forgery detected image (4.2), and finally, the performance is compared and shown as comparison graphs (4.3).

4.1 Result analysis

Various metrics are used to evaluate the performance of the proposed algorithm, which is described below:

4.1.1 True negative rate

It indicates the negative event rates and it is also called as specificity. The rate of the identified forged region is described as TNR, the value closer to one is better on this metric. The TNR is calculated by,

$$TNR = \frac{TN}{TN + FN} \tag{20}$$

4.1.2 False negative rate

It indicates the missing detection rate of the forged region. i.e., the detection system failed to detect the forged region is shown as FNR. Getting lower values on this metric is much better for improving the performance. The FNR is calculated by,

$$FNR = \frac{FN}{FN + TP} \tag{21}$$

4.1.3 False positive rate

It indicates the original images which are not detected as not forged. The FPR rate is calculated by the eq. 22,

$$FPR = \frac{FP}{FP + TN} \tag{22}$$

4.1.4 Precision

The precision rate indicates that the detected forged region is true. It is calculated by using the formula given in eq. 23:

$$\text{Pr}ecision = \frac{TP}{TP + FP} \tag{23}$$

4.1.5 Recall

The recall rate describes that the forged region is detected. The recall is also said to as True Positive Rate (TPR). The recall is calculated by,

$$\text{Re}call = \frac{TP}{TP + FN} \tag{24}$$

4.1.6 F1-score

It integrates the precision and recall rates and gives the single value. This integration is done by the eq. 25,

$$F_1 = 2\left(\frac{precision.recall}{precision + recall}\right) \tag{25}$$

4.1.7 Accuracy

The rate of accuracy indicates the correctly detected forged images among all the images present in the dataset. It is calculated by,

$$Acc = \frac{TP + TN}{FP + FN + FP + TN} \tag{26}$$

4.1.8 E-measure

The E-measure (Enhanced Alignment Measure) [10] shows the rate of foreground map on forged region of images and it is given by,

$$\text{E}-\text{measure} = \frac{1}{\text{wi} \times \text{hi}} \sum_{\text{x}=1}^{\text{wi}} \sum_{\text{y}=1}^{\text{hi}} \varphi\text{FM}(\text{x}, \text{y}) \tag{27}$$

In eq. 27, the width and height of map is given as wi and hi and the properties like image level statistics and pixel level matching is determined with alignment matrix φof binary map.

4.2 Various attacks in forged image

This section describes some of the attacks that are used for tampering the image and detection results of such attacks using the proposed method. Some of the most popular

image tampering attacks are the rotation, scaling, and noising. These attacks are efficiently detected with our extracted multi-scale feature using VGGNet. These features are extracted from various scales. So, it efficiently describes the matched original and tampered region with attacks like rotation and scaling. Various features from the multiple filters can describe the noised or low contrast region present in the image. The results of various attacked images are shown in Figs. 5, 6, and 7.

4.2.1 Rotation based attack

In this type of attack, the attackers rotate the copied objects in different angles. i.e., the copied region is rotated while pasting it into another region. Such rotation based attacks available in our dataset are in between 20 to 90. The proposed method successfully detected such rotation based forged images, as in Fig. 5:

4.2.2 Scaling based attack

These type of attacks will enlarge or decrease the size of the copied objects. The attackers maximize the height and width of the copied objects while pasting it to another region of the image. Scaling based forged images available in our dataset are scaled in between the range of 100 to 120. i.e., the forged objects are scaled up to 20% from original objects. These types of attacks are also detected successfully by the proposed method and its result is shown in Fig. 6:

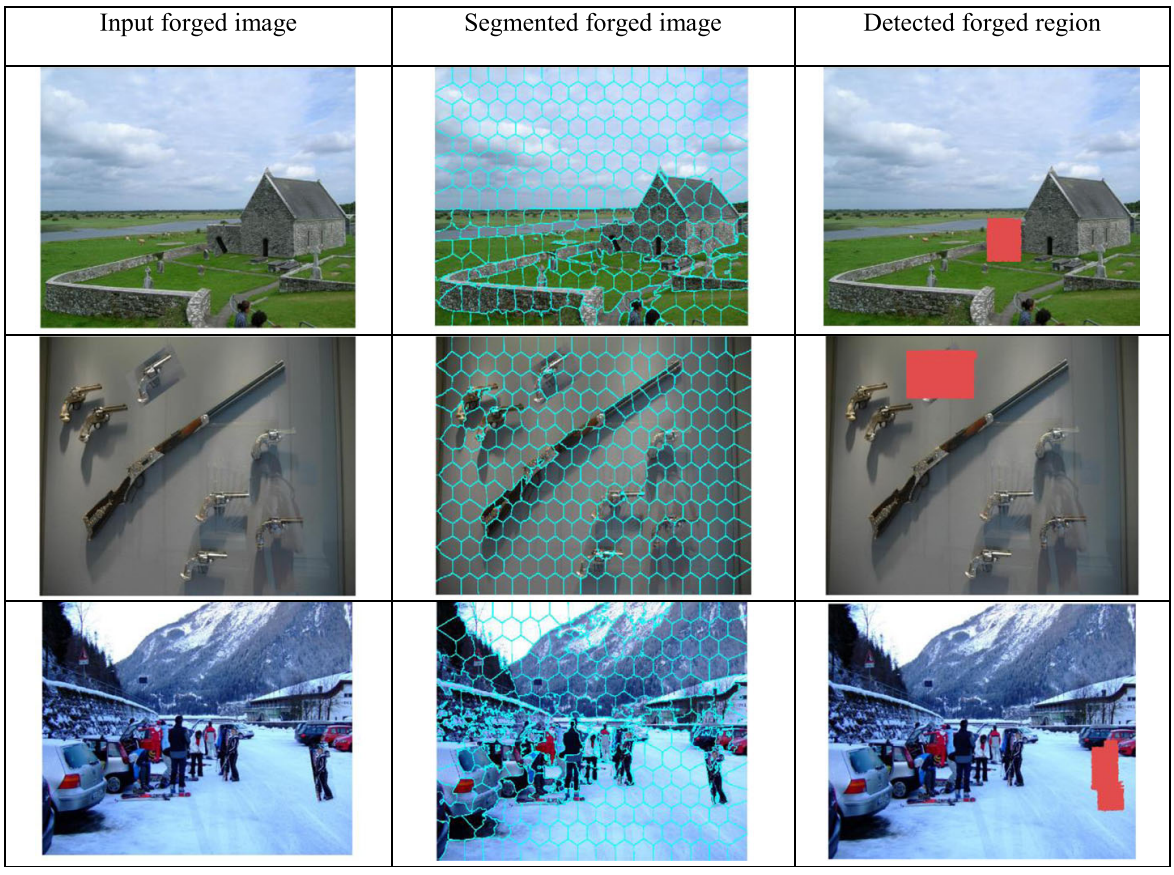


Fig. 5 Results of image forged based on rotation attack

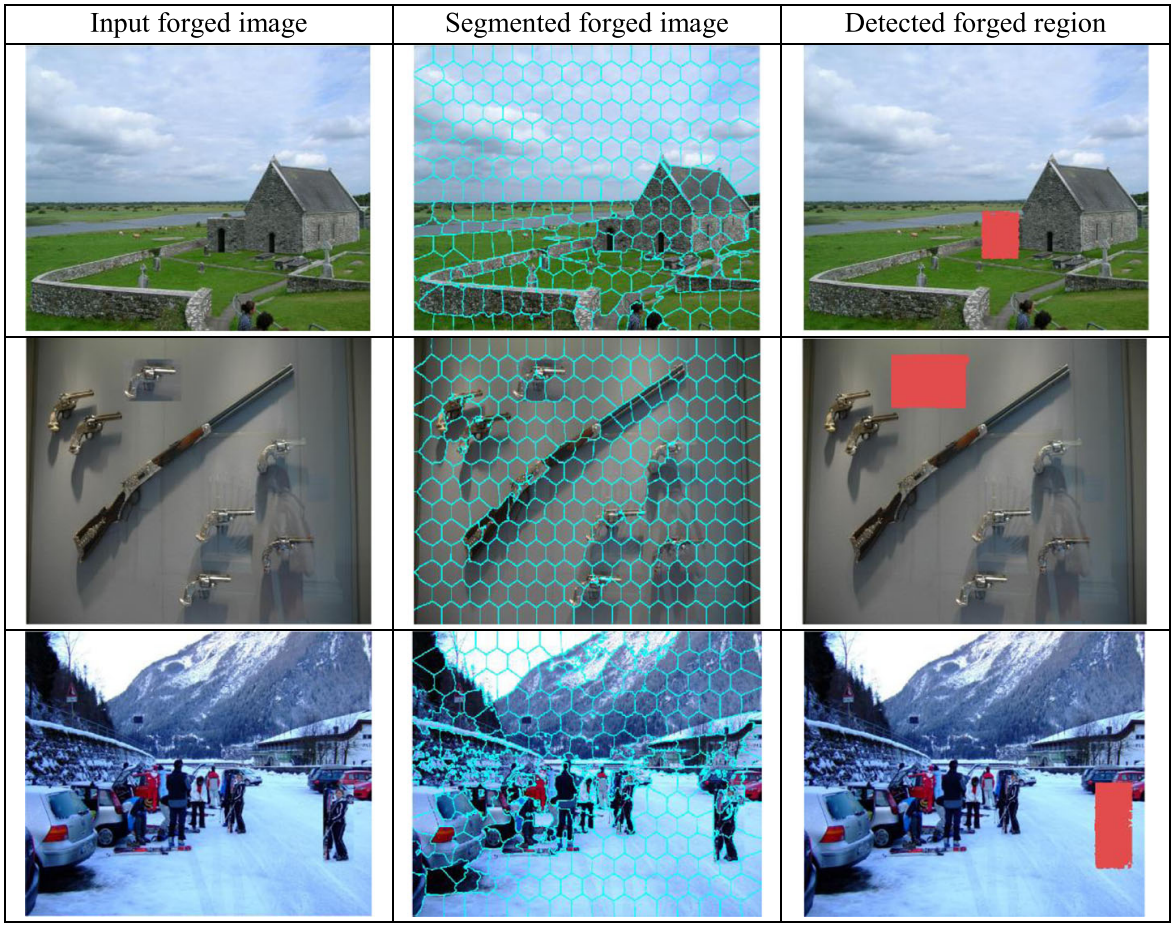


Fig. 6 Results of image forged based on scaling attack

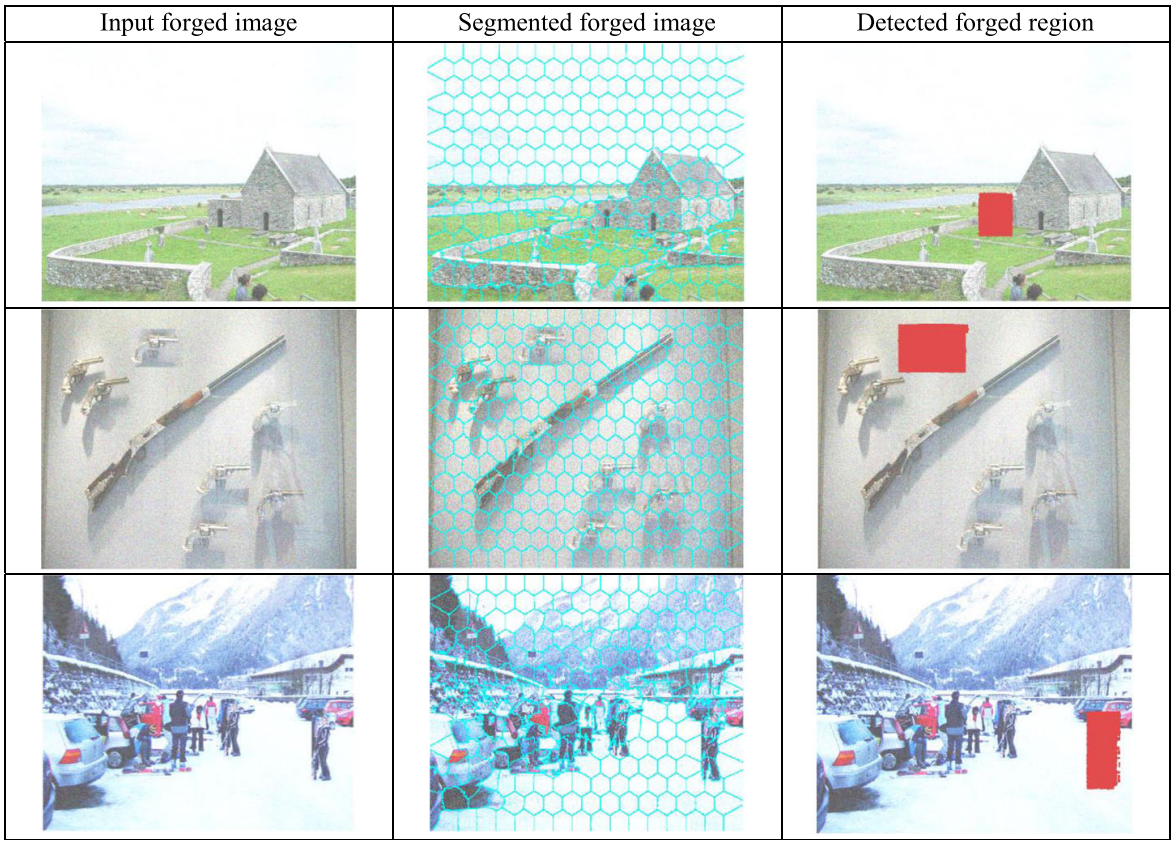


Fig. 7 Results of image forged based on noise attack

4.2.3 Noise based attack

In this type of attack, the copied region is normalized and add some Gaussian noise with the standard deviation. In our dataset, such noise based attacked images are normalized in the range of 0 to 1 and a standard deviation of 0.02, 0.04, 0.06, 0.08, 0.10. The result of this attack is shown in the Fig. 7:

In the Figs. 5, 6 and 7, the input and output of forged images in various transformation attacks are described. In the 1st column of Figs. 5, 6 and 7, the input tampered image is shown, in the 2nd column, the segmented patches of input images, and in 3rd column, the detected forged regions are marked (highlighted in red).

4.3 Performance graphs

This section describes the proposed method's performance as graphs using four existing methods [2, 14, 24, 26] for the performance comparison.

4.3.1 True negative rate

Figure 8 shows that the proposed system has higher TNR than other all existing methods. Here VGGNet has maximum TNR of 0.971.

4.3.2 False negative rate

Figure 9 describes that the VGGNet has lower FNR of 0.092. It is relatively low when comparing with other state-of-art methods.

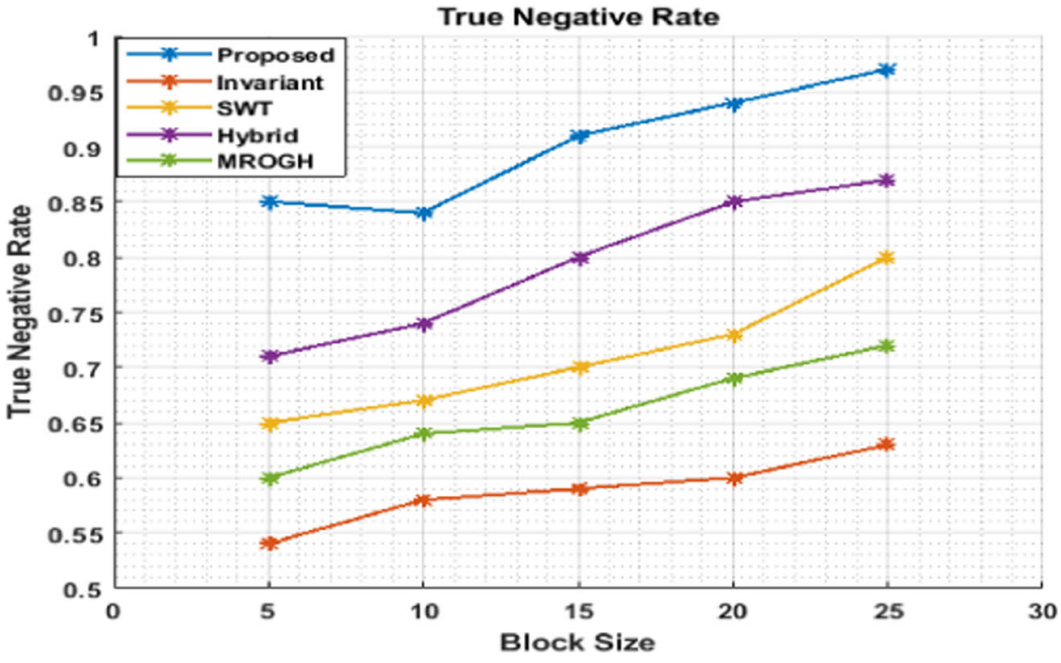


Fig. 8 TNR comparison

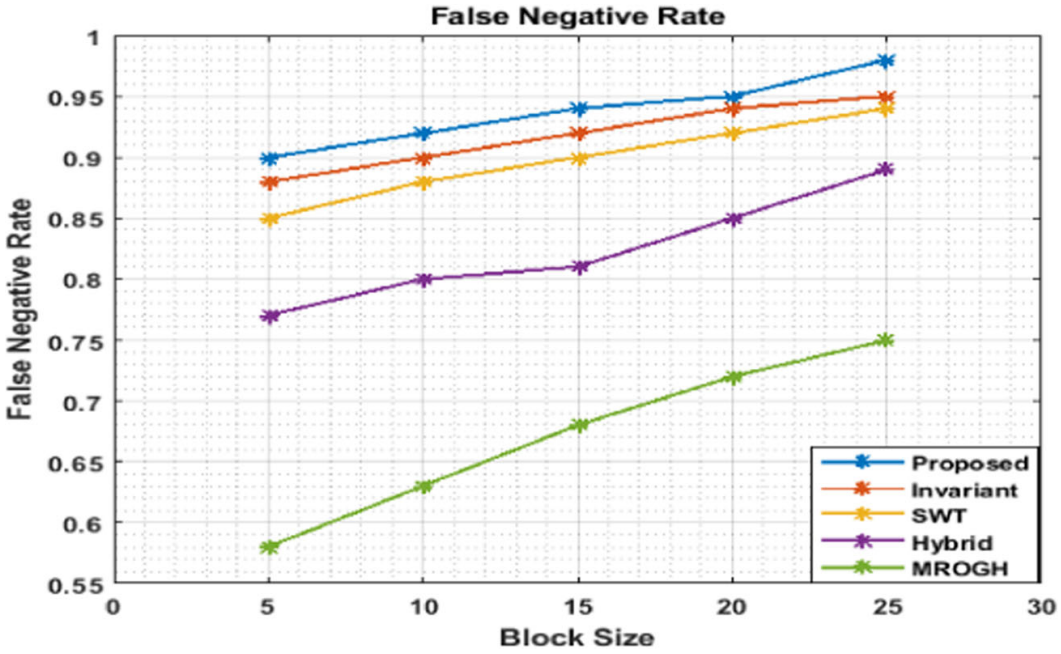


Fig. 9 FNR comparison

4.3.3 False positive rate

Better FPR is also achieved by the proposed VGGNet based method. Here, VGGNet has minimum of 0.55, invariant, SWT, hybrid and MROGH based method have achieved 0.62, 0.72, 0.65, and 0.75 as shown in Fig. 10.

4.3.4 Precision

Figure 11 describes the precision rate of proposed and existing methods. VGGNet achieved the precision rate of 98.026 as maximum.

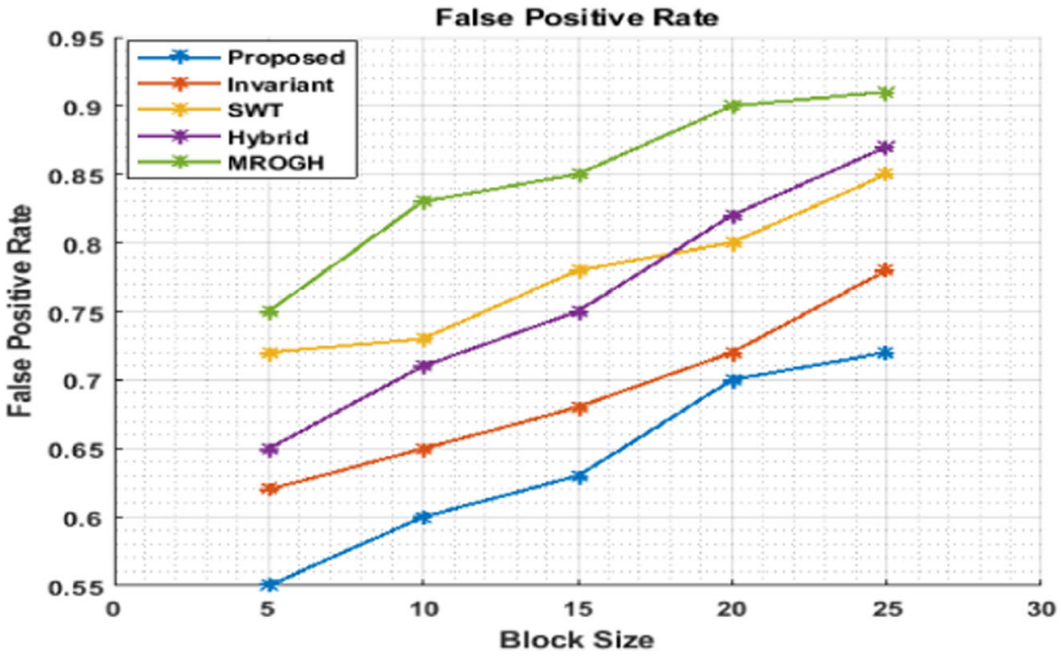


Fig. 10 FPR comparison

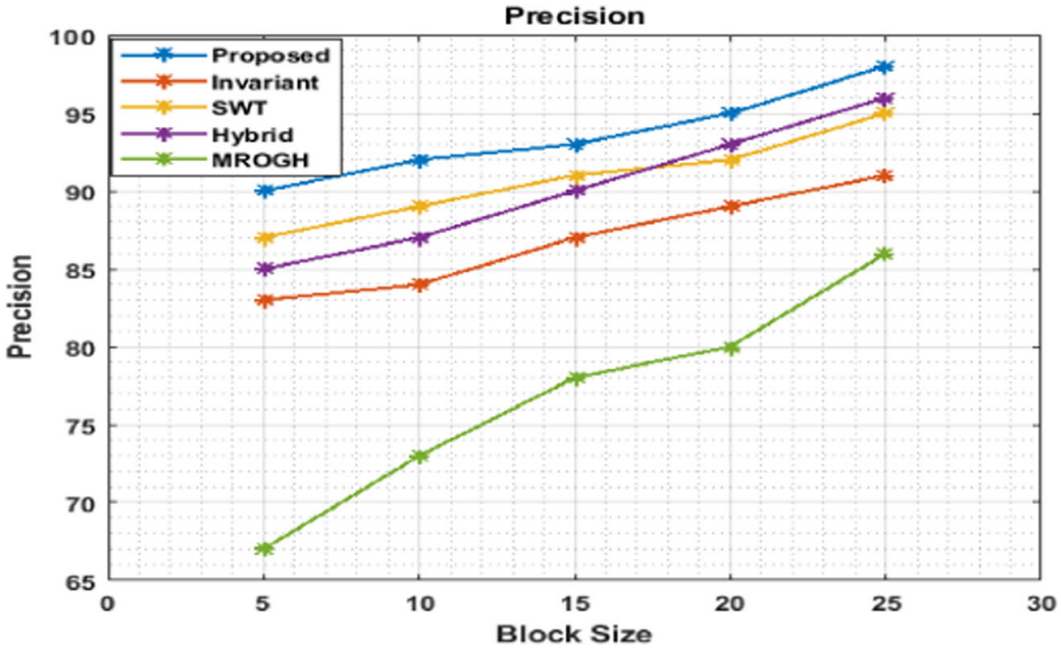


Fig. 11 Precision rate comparison

4.3.5 Recall

Figure 12 shows the recall rates, the existing methods has lower rates when comparing with VGGNet. It earned maximum rate 89.583 at recall.

4.3.6 F1-score

The F1-score of VGGNet is slight better than the existing methods, which is described in Fig. 13.

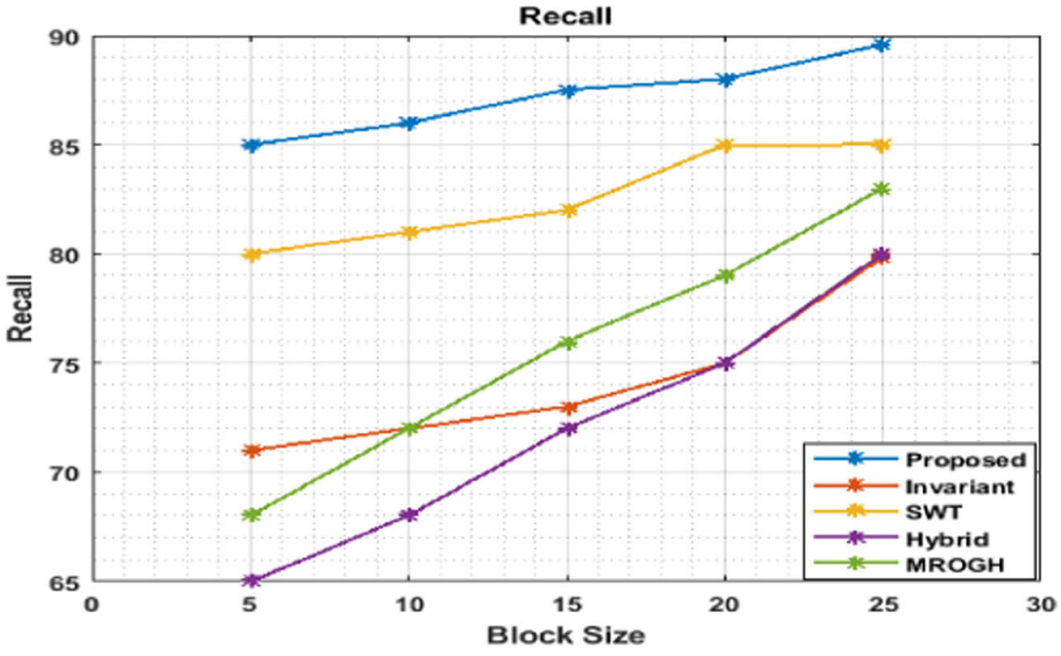


Fig. 12 Recall rate comparison

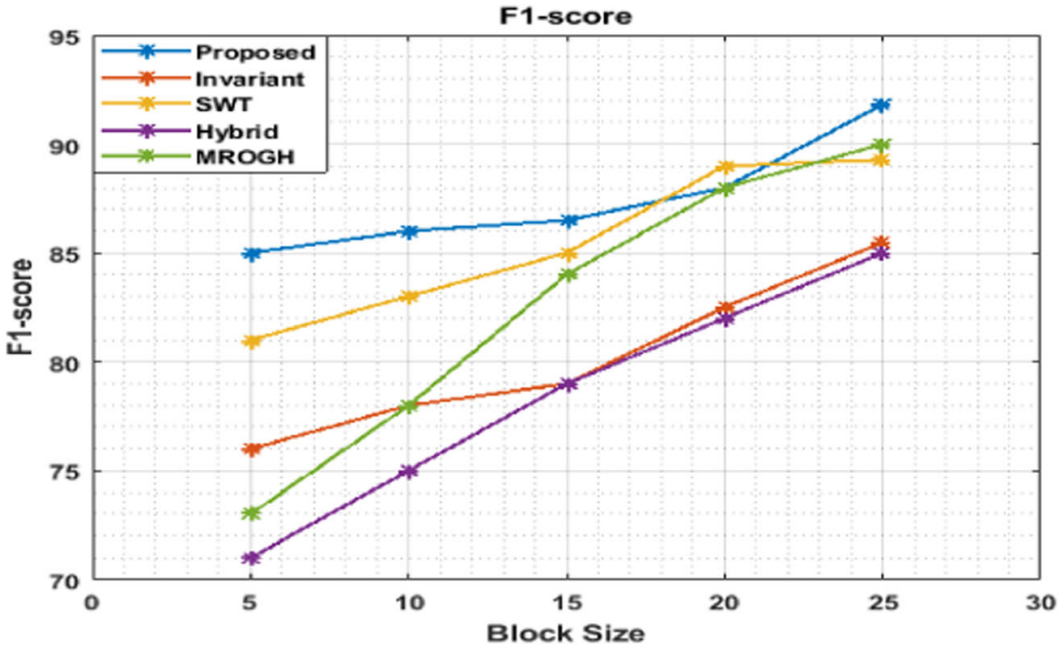


Fig. 13 F1-Score comparison

4.3.7 Accuracy

The VGGNet achieved good results on accuracy, the existing methods has 79%, 86%, 79%, and 83% accuracy rate. VGGNet earned maximum of 95% accuracy, it is much better when comparing with the existing methods and it is shown in Fig. 14.

4.3.8 E-measure

The comparison result of E-measure is described in Fig. 15, it clearly showing the higher result achieved by the proposed approach than all other existing approaches.

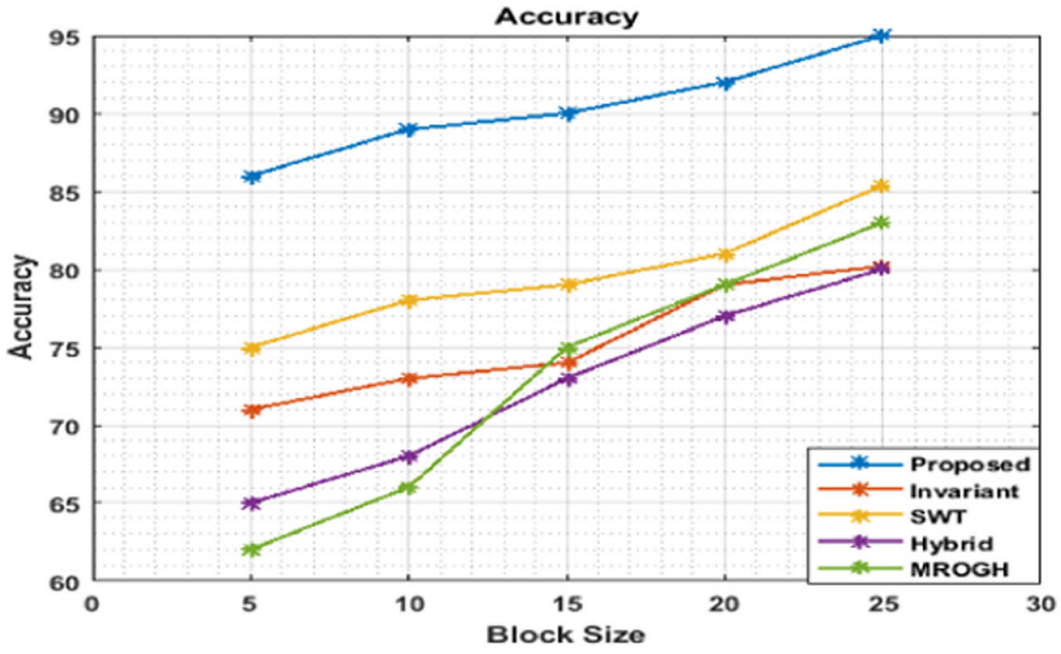


Fig. 14 Accuracy comparison

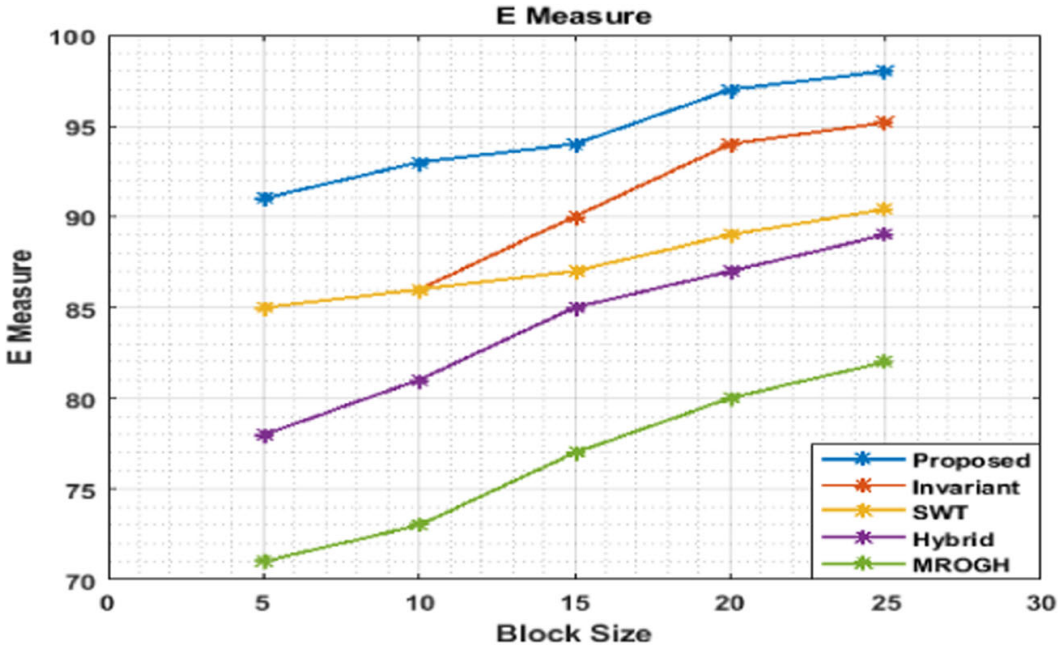


Fig. 15 E-measure comparison

The figures (Figs. 8, 9, 10, 11, 12, 13, 14 and 15), describes the performance of our system in terms of TNR, FNR, FPR, Precision, Recall, F1-Score, Accuracy, and E-measure. The performance of our proposed approach is improved when compared to another state of art methods [2, 14]. The proposed VGGNet based multi-scale feature extraction approach has better performance and the results show that the method is much efficient to handle various image forgery attacks such as rotation (Fig. 5), scaling (Fig. 6), noise (Fig. 7). Finally, it is shown that the proposed approach is better than other techniques by estimating several metrics and compared with two existing methods (Invariant feature-based method [2], SWT feature based method [14], hybrid feature based method [24], and MROGH feature based method [26]). However, the proposed approach does not detect the image forged using Multi-cloned attack. The patch matching procedure in proposed approach gets confused while matching multiple tampered patches in an image. It is a weakness of this approach and the advantage is predicting the forged region with other three types of attacks better than the existing ones.

5 Conclusion

Forgeries in the images are increasing day-by-day, the copy-move image forgery method is most popular and it can easily forge any image. Additionally, with these image forgery methods, some transformation attacks are also carried out for making the forged image undetectable by the forgery detection systems. So, it is much harder to detect the forged regions from the tampered images. In this article, the authors propose a method for detecting such image forgery based on the features extracted by using deep learning technique. In this method, the forged image is initialized as the input and then it is segmented using Simple Linear Iterative Clustering (SLIC). After the segmentation process, the features from the segmented patches are extracted. VGGNet has been used to extracts features from the various scales which are then used to reconstruct the dense depth of the image pixel. It helps to make the forged region to match with the original region. After the reconstruction process, the

adaptive patch matching (ADM) technique is applied to achieve the matched regions. Most of the suspicious regions will be visible at the end of this process. The achieved suspicious regions are merged with the segmented tampered image. During this process, the unforged regions are removed and the real forged regions are visible. The proposed approach is compared with some other existing techniques and results show that the proposed approach is better than the existing systems and effective even if the image is tampered by using various attacks.

References

1. Al-Qershi OM, Khoo BE (2017) Comparison of matching methods for copy-move image forgery detection. In 9th international conference on robotic, vision, signal processing and power applications (pp. 209-218). Springer, Singapore
2. Al-Qershi OM, Khoo BE (2018) Enhanced block-based copy-move forgery detection using k-means clustering', Multidimensional Systems and Signal Processing, pp. 1–25
3. Ansari MD, Ghrera SP (2018) Copy-move image forgery detection using direct fuzzy transform and ring projection. International Journal of Signal and Imaging Systems Engineering 11(1):44–51
4. Bi X, Pun CM (2017) Fast reflective offset-guided searching method for copy-move forgery detection. Inf Sci 418:531–545
5. Bi X, Pun CM (2018) Fast copy-move forgery detection using local bidirectional coherency error refinement. Pattern Recogn 81:161–175
6. Bi X, Pun CM, Yuan XC (2018) Multi-scale feature extraction and adaptive matching for copy-move forgery detection. Multimed Tools Appl 77(1):363–385
7. Bi XL, Pun C-M, Yuan X-C (2018) Multi-scale feature extraction and adaptive matching for copy-move forgery detection. Multimed Tools Appl 77(1):363–385
8. Chou CL, Lee JC (2017) Copy-move forgery detection based on local Gabor wavelets patterns. In international conference on security with intelligent computing and big-data services (pp. 47-56). Springer, Cham
9. Christlein V, Riess C, Jordan J, Riess C, Angelopoulou E (2012) An evaluation of popular copy-move forgery detection approaches. IEEE Transactions on Information Forensics and Security 7(6):1841–1854
10. Fan D-P, Gong C, Cao Y, Ren B, Cheng M-M, Borji A (2018) Enhanced-alignment measure for binary foreground map evaluation. Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18), pp.698-704
11. Hammad R (2017) Blind copula based copy-move forgery detection algorithm. IEEE international conference on consumer electronics (ICCE) (pp. 436-437). IEEE
12. Ma J, Wang X, Xiao B (2015) An image segmentation method based on Simple Linear Iterative Clustering and graph-based semi-supervised learning. In Orange Technologies (ICOT), 2015 International Conference on, pp. 10–13. IEEE
13. Mahmood A, Bennamoun M, An S, Sohel F, Boussaid F, Hovey R, Kendrick G, Fisher RB (2016) Coral classification with hybrid feature representations. In Image Processing (ICIP), 2016 IEEE International Conference on, pp. 519–523. IEEE
14. Mahmood T, Mehmood Z, Shah M, Saba T (2018) A robust technique for copy-move forgery detection and localization in digital images via stationary wavelet and discrete cosine transform. J Vis Commun Image Represent 53:202–214
15. Manu VT, Mehtre BM (2018) Copy-move tampering detection using affine transformation property preservation on clustered keypoints. SIViP 12(3):549–556
16. Novozámský A, Šorel M (2018) Detection of copy-move image modification using JPEG compression model. Forensic Sci Int 283:47–57
17. Oyiza AH, Maarof MA (2018) An improved DCT block based technique for copy-move forgery detection in medical images. International Journal of Innovative Computing 8(1)
18. Panda S, Mishra M (2018) Passive techniques of digital image forgery detection: developments and challenges. In advances in electronics, communication and computing (pp. 281-290). Springer, Singapore
19. Soni B, Das PK, Thounaojam DM (2018) multiCMFD: fast and efficient system for multiple copy-move forgeries detection in image. In proceedings of the 2018 international conference on image and graphics processing (pp. 53-58). ACM

20. Wang XY, Liu YN, Xu H, Wang P, Yang HY (2018) Robust copy–move forgery detection using quaternion exponent moments. Pattern Anal Applic 21(2):451–467
21. Warif NBA, Wahab AWA, Idris MYI, Ramli R, Salleh R, Shamshirband S, Choo KKR (2016) Copy-move forgery detection: survey, challenges and future directions. J Netw Comput Appl 75:259–278
22. Wu Y, Abd-Almageed W, Natarajan P (2018) Image copy-move forgery detection via an end-to-end deep neural network. In applications of computer vision (WACV), 2018 IEEE winter conference on (pp. 1907-1915). IEEE
23. Yang F, Li J, Lu W, Weng J (2017) Copy-move forgery detection based on hybrid features. Eng Appl Artif Intell 59:73–83
24. Yang F, Li J, Lu W, Weng J (2017) Copy-move forgery detection based on hybrid features. Eng Appl Artif Intell 59:73–83
25. Yang B, Sun X, Guo H, Xia Z, Chen X (2018) A copy-move forgery detection method based on CMFD-SIFT. Multimed Tools Appl 77(1):837–855
26. Yu L, Han Q, Niu X (2016) Feature point-based copy-move forgery detection: covering the non-textured areas. Multimed Tools Appl 75(2):1159–1176
27. Zhong J, Gan Y, Young J, Huang L, Lin P (2017) A new block-based method for copy move forgery detection under image geometric transforms. Multimed Tools Appl 76(13):14887–14903

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Ritu Agarwal is working as an Assistant Professor in the Department of Information Technology at Delhi Technological University, Delhi India and currently pursuing her Ph.D. Her research interests include Information Security, Data Recovery, Digital Media Forensics, Cyber Forensics.



Om Prakash Verma received the B.E. degree from the Malaviya National Institute of Technology, Jaipur, India, the M.Tech. degree from the Indian Institute of Technology, Delhi, New Delhi, India, and the Ph.D. degree in the area of applications of soft and evolutionary computing in image processing from the University of Delhi, New Delhi, India. From 1992 to 1998, he was an Assistant Professor with Department of Electronics and Communication Engineering, Malaviya National Institute of Technology, Jaipur, India. He joined the Department of Electronics and Communication Engineering, Delhi Technological University (formerly Delhi College of Engineering), New Delhi, India, as an Associate Professor in 1998, and is currently Professor at Delhi Technological University . His current research interests include applied soft computing, nature inspired algorithms, swarm intelligence, evolutionary computing, image processing and digital forensics.

Microchimica Acta (2020) 187:1
https://doi.org/10.1007/s00604-019-3921-8

ORIGINAL PAPER

An impedimetric biosensor based on electrophoretically assembled ZnO nanorods and carboxylated graphene nanoflakes on an indium tin oxide electrode for detection of the DNA of *Escherichia coli* O157:H7

Nandita Jaiswal[1] · Chandra Mouli Pandey[2] · Shipra Solanki[3] · Ida Tiwari[1] · Bansi Dhar Malhotra[3]

Received: 23 May 2019 / Accepted: 11 October 2019
© Springer-Verlag GmbH Austria, part of Springer Nature 2019

Abstract
Aminopropyltrimethoxysilane (APTMS)-functionalized zinc oxide (ZnO) nanorods and carboxylated graphene nanoflakes (c-GNF) were used in a composite that was electrophoretically deposited on an indium tin oxide (ITO) coated glass substrate. The modified ITO electrodes were characterized using various microscopic and spectroscopic techniques which confirm the deposition of the APTMS-ZnO/c-GNF composite. The electrodes have been used for the covalent immobilization of an *Escherichia coli* O157:H7 (*E. coli*)-specific DNA prob. Impedimetric studies revealed that the gene sensor displays linear response in a wide range of target DNA concentration (10^{-16} M to 10^{-6} M) with a detection limit of 0.1 fM. The studies on the cross-reactivity to other water-borne pathogens show that the bioelectrode is highly specific.

Keywords Biosensor · *Escherichia coli* · Zinc oxide · Graphene nanoflakes · Electrophoretic deposition

Introduction

Nanostructured materials have gathered interest as functional materials in various aspects of designing biomedical devices [1–5]. Nanostructured metal oxides (NMOs) have been widely used in biosensing applications due to their various characteristics, such as large surface to volume ratio, high electrocatalytic activity, their small band gap that enhances their electron conductivity and the adsorption ability [6] which helps in increasing the loading of biomolecules per unit mass of particles. The employment of NMOs may open the possibility of new signal transduction due to their sub-micrometer dimension and can be useful for in vivo analysis. The NMOs reported to have their potential application in the field of biosensor are of zinc, iron, cerium, tin, zirconium, titanium and magnesium [7]. Among these zinc oxide (ZnO), has been widely used for the development of various sensing platforms. ZnO has gained much interest due to its novel characteristics such as biocompatibility, catalytic efficiency, non-toxicity, as well as chemical and electrical stability [8–11]. ZnO shows various nanostructures of different morphologies such as nanowires, nanoflakes, nanorods, nanobelts, nanoneedles and nanocombs [12–16].

Though NMOs have shown excellent properties as materials for biosensing fabrication, but there are some limitations such as low conductivity as compared to carbon materials, low surface area and high resistance. So, it will be logical to develop a composite material containing both carbon and transition metal oxide as the electrode material, combining the merits and mitigating the shortcomings of both the individual components [17, 18]. The use of graphene in nanostructured carbon-metal oxides composites not only provides physical support to metal oxides but also serve as channels for charge transport due to its high surface area and the robust conductive

Nandita Jaiswal and Chandra Mouli Pandey contributed equally to this work.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00604-019-3921-8) contains supplementary material, which is available to authorized users.

✉ Ida Tiwari
idatiwari@bhu.ac.in

[1] Department of Chemistry(Centre of Advanced Study), Institute of Science, Banaras Hindu University, Varanasi 221005, India

[2] Department of Applied Chemistry, Delhi Technological University, Delhi 110042, India

[3] Department of Biotechnology, Delhi Technological University, Delhi 110042, India

Published online: 03 December 2019

structure. In graphene-metal oxide composite, the graphene provides the benefit of high electron conductivity while electro-activities of metal oxides contribute to high specific capacitance. The performance of biosensors based on such materials is governed by the interface between the nanostructured composite material and biomolecule [19]. The nano bio-interface formation and its properties depend upon nature of the nanostructured material and also other parameters related to effective surface area, roughness, porosity, conductance, functional groups etc. The biomolecules are said to bind the metal nanocomposites generally by physical adsorption or chemical binding [20] and the activity of the biomolecule can be retained by establishing a biocompatible microenvironment on the biointerface [21, 22]. Thus the use of nanostructured composite material would decide the sensitivity, selectivity, detection limit, cost and shelf-life of the biosensor.

The enter-hemorrhagic stereotype of bacterium *Escherichia coli* is *E. coli* O157:H7 and its infection can lead to kidney failure and hemorrhagic diarrhea [23]. Various methods such as culture and colony counting are available for the detection of the pathogen. The traditional methods require an excess of time and the results are also ambiguous [24]. Some methods involve enrichment of culture by growing it for a specific time and then separated and identified using polymerase chain reaction or immunological methods of identifying genes. These methods are quite complicated and costly in terms of requirement of labeled antibodies and their pre-separation or pre-enrichment. Thus, there is an urgent requirement for a rapid, sensitive and cost-effective technique to overcome these problems. An alternative technique to the methods mentioned above is the use of biosensors for pathogenic bacterial genome detection [25, 26].

The biosensors based on nanostructured composite materials for nucleic acid detection have aroused much more interest due to their efficiency of converting the hybridization event into an analytical signal. Wang et al. reported a biosensor based on ZnO nanowires, multiwalled carbon nanotubes (MWCNT) and gold nanoparticles composite utilizing thiolated single-stranded DNA probe to immobilize it on the gold nanoparticle surface. The detection limit of this DNA sensor was found to be 0.35 pM using $[Ru(NH_3)_6]^{3+}$ as a redox indicator [27]. Sol-gel derived nanostructured ZrO_2 film for the detection of 16 s rRNA coding region for *E.coli* has been reported with a sensitive determination in the range of 10^{-6} to 10^{6} pM of complementary DNA [28]. Yang et al. reported a composite based on nano ZrO_2, Chitosan and MWCNT for the development of DNA sensor on glassy carbon electrode. The composite provided good charge transport characteristics with high surface area favoring the increase in loading capacity and enhanced selectivity with a detection limit of 75 pM [29]. For point of care analysis, an ideal detection method should be such that it may detect the disease directly from food and water samples with minimal requirement of sample preparation. However, the drawbacks with nucleic acid based biosensor are that the extraction of DNA from culture samples and processing involve laborious pre–treatment steps. Thus efforts are being made to develop affinity based biosensors (using antibody–antigen interactions) and aptasensors to allow the detection of *E. coli* O157:H7 cells directly from food and water samples but till date they are not very applicable and have not been fully developed into technology. Hence, it becomes imperative to design such materials which can overcome these drawbacks.

The present work reports studies conducted to exploit the properties of ZnO nanorods by combining it with a new class of graphene material known as carboxylated graphene nanoflakes (c-GNF) [30]. The ZnO nanorods were firstly modified with aminopropyltrimethoxysilane (APTMS) to introduce NH_2 groups on the surface of ZnO nanorods. Then the electrostatic interaction of this APTMS functionalized ZnO nanorods (APTMS-ZnO) with the carboxylic acid groups on c-GNF was done for the formation of APTMS functionalized ZnO nanorods-carboxylated graphene nanoflakes composite (APTMS-ZnO/c-GNF) composite. Further, the composite was electrophoretically deposited using magnesium (II) electrolyte which provided overall positive charge to the colloid suspension for the fabrication of impedimetric biosensor for *E. coli* DNA detection.

Chemicals and materials required

Zinc(II) nitrate hexahydrate (www.alfa.com), aminopropyltrimethoxysilane (www.sigmaaldrich.com), glutaraldehyde (www.spectrochem.co.in) and sodium hydroxide pellets (www.sdfine.com). The specific sequence of *E. coli* probe (17 mer.) has been identified from the 16 s rRNA coding region of the *E. coli* genome. The complementary, non-complementary and one-base mismatch target sequences were procured from Sigma Aldrich (www.sigmaaldrich.com). Sequences of oligonucleotide probes used are:

Probe I- probe DNA (pDNA): amino-5′-GGT CCG CTT GCT CTC GC-3′.
Probe II- complementary DNA (cDNA): 5’-GCG AGA GCA AGC GGA CC-3′.
Probe III- non-complementary DNA: 5’-CTA GTC GTA TAG TAG GC-3′.
Probe IV- one-base mismatch DNA: 5’-GCG AGA G**A**A AGC GGA CC-3′.

The solution of oligonucleotide was made in Tris–EDTA buffer (1 M Tris–HCl, 0.5 M EDTA, pH 8.0) (www.merck.com) and was stored at −20° C prior to use.

Apparatus and measurements

X-ray diffraction (XRD) analysis was performed using D8 Advance/Discover Bruker, Germany, Diffractometer with Cu Kα. The UV-vis spectra was recorded using Hitachi U-3900 spectrophotometer while the transmission electron microscopy (TEM) and energy dispersive X-ray analysis (EDX) were carried out using TECNAI 12 G2FEI microscope, operating at 120 kV instrument. FT-IR spectroscopic study was done using PerkinElmer spectrometer (Spectrum two, UK), the scanning electron microscopy (SEM) images were recorded using a JEOLJSM-6700F field-emitting scanning electron microscope (FESEM, 15 kV) while the contact angle was measured using contact angle meter (Data physics OCA15EC). Electrochemical characterization were carried out on an Autolab potentiostat/galvanostat (www.metrohm-autolab.com) using a three-electrode system with ITO as working electrode, Ag/AgCl as a reference electrode, and platinum foil as a counter electrode in phosphate buffer solution (PBS of 100 mM concentration with pH 7.0 in 0.9% NaCl).

Experimental section

Synthesis of APTMS functionalized ZnO nanorods

The synthesis of ZnO nanorods are discussed in section 1 of supp. Material [31]. For functionalization of the ZnO nanorods, APTMS was used. For amino functionalization of ZnO, 0.2 g ZnO powder was ultrasonicated in 20 mL toluene followed by addition of 400 μL of APTMS under nitrogen atmosphere in a round bottom flask and refluxed at 110 °C for 6 h to get APTMS functionalized ZnO nanorods (APTMS-ZnO) [32, 33].

Synthesis of APTMS functionalized ZnO nanorods-carboxylated graphene nanoflakes composite

The c-GNF were synthesized as reported in the literature [30]. A dispersion of c-GNF was made using DI water (0.5 $mgmL^{-1}$) via ultrasonication for 2 h. Further 10 mg of APTMS-ZnO nanorods were added to the above dispersion and ultrasonicated for 2 h followed by subsequent stirring (4 h). Then the suspension was centrifuged at 10,000 rpm for 1 h to obtain APTMS-ZnO nanorods-c-GNF composite (APTMS-ZnO/c-GNF). Finally, the obtained material was washed with distilled water and then dried at 70 °C under vacuum (Scheme 1a).

Electrophoretic deposition of APTMS-ZnO/c-GNF

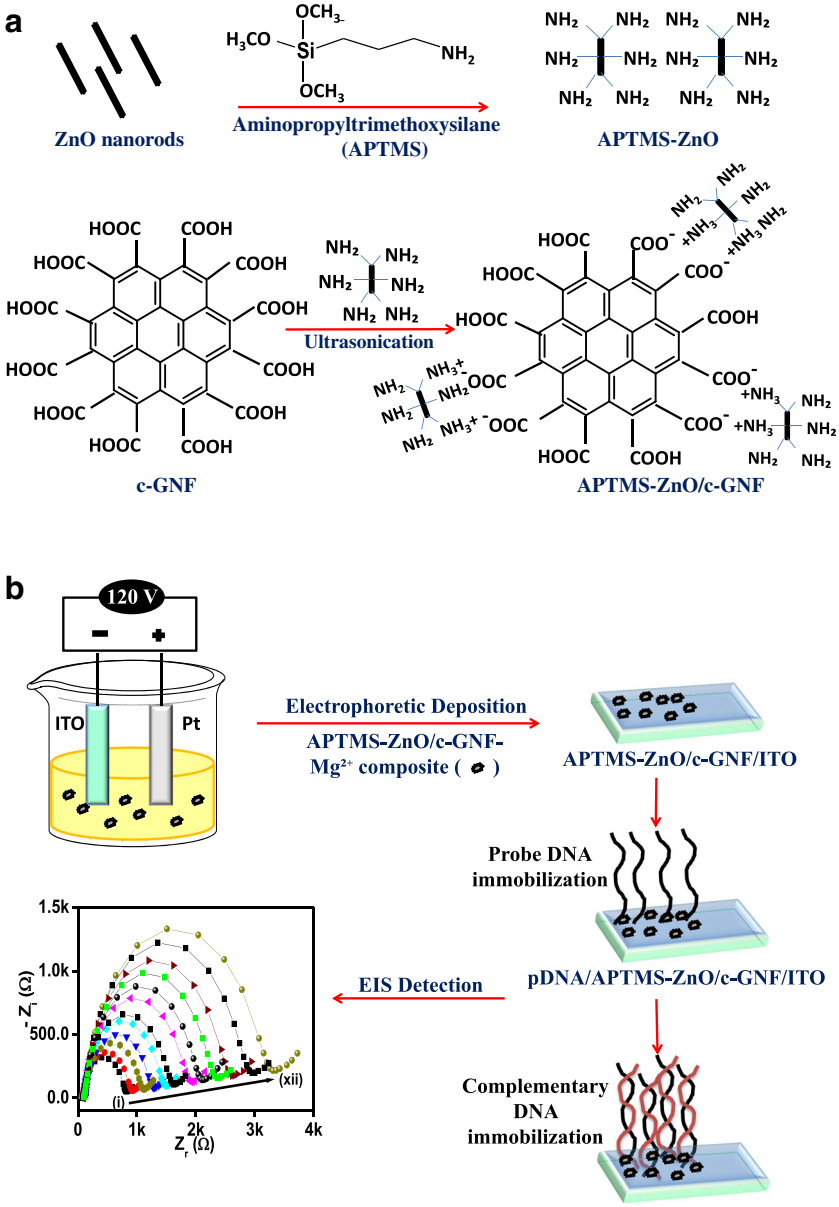
Prior to deposition, the ITO coated glass electrode were hydrolyzed by immersing in $H_2O_2/NH_4OH/H_2O$ solution (1:1:5, v/v) for about 30 min at 80 °C. For the deposition, APTMS-ZnO/c-GNF (5 mg) were dispersed in 10 mL acetonitrile. The electrophoretic deposition was carried using the DC source by applying a voltage of 120 V for 2 min. 10^{-4} M of $Mg(NO_3)_2.6H_2O$ was added to the colloidal suspension (APTMS-ZnO/c-GNF in acetonitrile) acting as an electrolyte (for creating a charge on the colloids).

Immobilization and hybridization of probe DNA on APTMS-ZnO/c-GNF/ITO electrode

The coupling of amine terminal of probe DNA (pDNA) and APTMS-ZnO/c-GNF/ITO electrode was done by immobilizing 20 μL of pDNA on the modified electrode using glutaraldehyde as cross linker (0.1%, v/v; 4 h). The fabricated pDNA/APTMS-ZnO/c-GNF/ITO bioelectrode was employed for the detection of *E. coli* complementary DNA (cDNA). The different concentrations of pDNA and cDNA were prepared in Tris-EDTA buffer (1 M Tris−HCl, 0.5 M EDTA) of pH 8.0 from the stock solution of 1×10^{-6} M concentration. The pDNA/APTMS-ZnO/c-GNF/ITO was subjected to different concentrations of cDNA for 30 min and the corresponding Rct value was measured using electrochemical impedance spectroscopy. The different steps involved in the fabrication of APTMS-ZnO/c-GNF/ITO and pDNA/APTMS-ZnO/c-GNF/ITO has been depicted in scheme 1b.

Extraction and pretreatment of DNA from bacterial clinical samples

The extraction of DNA was carried out using the panel of strains comprising *E. coli*, *Salmonella Typhimurium*, *Neisseria meningitides* and *Klebsiella pneumonia*. The DNA was extracted by suspending the bacterial colonies and vortexing and pouring them into 100 mL of MilliQ water. The suspension was further boiled for 10 min and centrifuged at 10,000 rpm for 5 min and equal volume (100 mL) of 24:1 (v/v) chloroform–iso-amyl alcohol was added to the boiled and separated suspension. This solution was again centrifuged at 12,000 rpm for 10 min, which resulted in the formation of an aqueous layer containing DNA. This aqueous layer was carefully separated out and kept at -20 °C prior to use. For denatured single-stranded DNA, the DNA samples extracted from the above-mentioned method were heated on water bath (95 °C) for 5 min and were immediately chilled in ice. These aliquots of samples were subjected to sonication for 15 min at 120 V to break the long DNA strands into smaller fragments.



Scheme 1 Schematic representation of **(a)** synthesis of APTMS-ZnO/c-GNF and **(b)** subsequent fabrication of nucleic acid biosensor for *E. coli* DNA detection using APTMS-ZnO/c-GNF/ITO electrode.

Results and discussion

Morphological characterization of the deposited materials

The morphology of the electrodeposited APTMS-ZnO/c-GNF on ITO electrode was studied using SEM (Fig. 1). The SEM image of APTMS-ZnO/c-GNF/ITO shows dendritic structure that might be due to the assembly of APTMS-ZnO nanorods with c-GNF on applying potential during the electrodeposition (Fig. 1a and b at different magnification). Even the rods of the ZnO have not lost its morphology and dendritic structure (Fig. 1c) would lead to a greater surface area of the APTMS-ZnO/c-GNF/ITO which will provide an excellent platform for pDNA attachment. While in the case of electrodeposition of APTMS-ZnO no such assembly is observed and smoothsheet-like morphology can be seen for APTMS-ZnO/ITO (Fig. 1d).

Electrochemical characterization

Electrochemical impedance spectroscopy

The electrochemical properties of the deposited material are evaluated using electrochemical impedance spectroscopic (EIS) studies since it is a powerful tool for observing the charge transfer processes at the electrode-electrolyte interface [25]. The impedimetric spectra are seen in Fig. 2a called as the

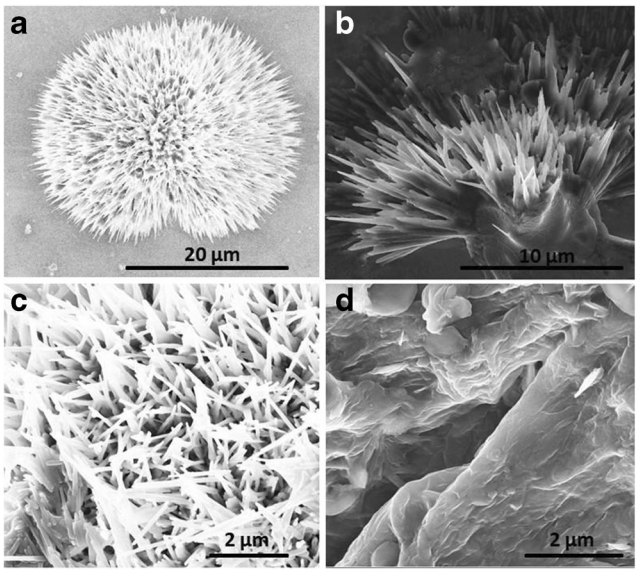


Fig. 1 (**a**), (**b**), (**c**) SEM images of APTMS-ZnO/c-GNF/ITO electrode at different magnifications and (**d**) SEM image of APTMS-ZnO/ITO electrode

Nyquist plot which consists of two parts one semicircular part at high frequency and the other linear plot at a lower frequency. The semicircular high-frequency part signifies the electron transfer mediated process. The linear lower frequency part corresponds to the diffusion-limited process. The impedimetric spectrum of the fabricated biosensor was fitted using a modified Randles equivalent circuit. This circuit is made up of different elements such as electrolyte resistance (Rs), Warburg impedance (Z_W), double layer capacitance (CPE) and the electron transfer resistance (Rct). The Rs is the resistance which is observed between the working and the reference electrode. Z_W is the element associated with the process at the electrode-electrolyte interface due to diffusion of redox probe ions. CPE is related to the double layer capacitance. The numerical value of Rct is helpful in deriving the electrochemical response of the fabricated biosensor. The measurements of the impedance were made to carry out at the open circuit voltage and the test frequency was in the range of 100 kHz to 0.1 Hz with amplitude of ±5 mV [25].

The EIS technique was adopted to investigate the electrochemical behavior of APTMS-ZnO/c-GNF/ITO electrode, pDNA/APTMS-ZnO/c-GNF/ITO electrode and pDNA/APTMS-ZnO/c-GNF/ITO electrode incubated with *E. coli* cDNA by comparing their Rct values (Fig. 2a). The semicircular region of the impedance curve reflects the numerical value of Rct, derived by fitting a Randles and Ershler model circuit. Also, the presence of a functional group on the composite material APTMS-ZnO/c-GNF might chemisorb the complementary target of the oligonucleotide. To avoid the false positive signal due to chemisorption, there is a need to optimize the concentration of pDNA. The optimized concentration of pDNA was found to be 10^{-6} M, which showed the leveling of the Rct value (optimization data not shown).

Cyclic voltammetry

The cyclic voltammetry studies were used to investigate the electrocatalytic behavior of the fabricated electrodes towards the redox behavior of $[Fe(CN)_6]^{3-/4-}$ [25].The outcome of the cyclic voltammetry as the function of scan rate (υ) for APTMS-ZnO/c-GNF/ITO and pDNA/APTMS-ZnO/c-GNF/ITO can be seen in Fig. 2b and c. The anodic and cathodic peak corresponds to oxidation and reduction of $[Fe(CN)_6]^{3-/4-}$ redox couple respectively. The anodic and cathodic peak currents were found to vary linearly with υ in the range of 10 to 300 mVs^{-1} indicating the electrochemical reaction of the redox pair at the working electrode to be diffusion-limited (Fig. S7 A and B).

Electrochemical response study

The electrochemical response of the DNA hybridization was studied using the EIS technique [34]. Upon immobilization of pDNA on APTMS-ZnO/c-GNF/ITO an increase in the Rct value (754 Ω) was observed because of the repulsion between the $[Fe(CN)_6]^{3-/4-}$ electrolytic redox couple and the phosphate backbone of the pDNA (Fig. 3a curve (i)). There was a further increase in Rct value after the hybridization of complementary DNA (889 Ω) due to an increase in the number of phosphate backbone in the double-stranded DNA (Fig. 3a curve (ii-xii)). It can be inferred that with a change in the negative charge and conformation transition, the interfacial properties changes and this becomes the basis of sensitive determination of cDNA of *E. coli*. The signal used for the measurement was the difference in the Rct value of pDNA immobilized electrode and the electrode after hybridization with cDNA {ΔRct = Rct(cDNA) – Rct(pDNA)}. There was a linear relationship between the analytical signal (ΔRct) and the log of concentration of cDNA target ranging from 10^{-16} to 10^{-6} M (Fig. 3b) and follows the following regression eq. (1):

$$\Delta\mathrm{Rct} = 3840.02 + 235.14\,[\log(\text{concentration of cDNA})] \quad (1)$$

with linear regression of 0.9952. The detection limit of the fabricated pDNA/APTMS-ZnO/c-GNF/ITO based biosensor is 0.1 fM. As the APTMS-ZnO/c-GNF contain ample functional groups which provide a conducive platform for the immobilization of pDNA, leading to higher loading of pDNA on to the APTMS-ZnO/c-GNF/ITO electrode, resulting in increased hybridization efficiency. From the data obtained it can be easily concluded that compared with other previously

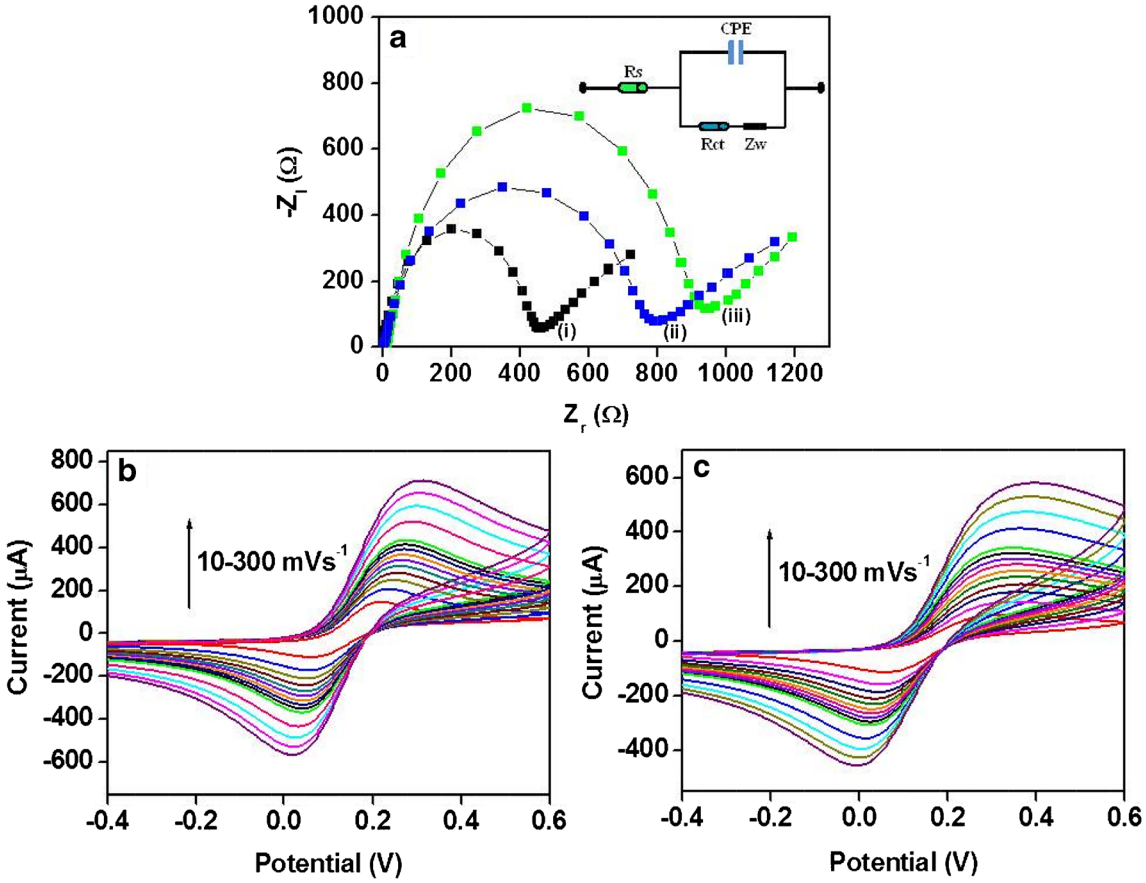


Fig. 2 (**a**) Nyquist diagram (Z_i vs. Z_r) for the Faradic impedance in the frequency range from 100 kHz to 0.1 Hz for (i) APTMS-ZnO/c-GNF/ITO electrode, (ii) pDNA/APTMS-ZnO/c-GNF/ITO bioelectrode, and (iii) cDNA on pDNA/APTMS-ZnO/c-GNF/ITO bioelectrode., (**b**) Cyclic voltammogram of APTMS-ZnO/c-GNF/ITO electrode and (c) pDNA/APTMS-ZnO/c-GNF/ITO electrode (in 5 mM $[Fe(CN)_6]^{3-/4-}$, PBS, 100 mM, pH 7.0, 0.9% NaCl)

reported methods which have their inherent limitations as discussed in preceding section 1 and other works reported based on hybridization this genosensor shows a wider detection range, better sensitivity and low detection limit (cf. Table S3 and S4).

Selectivity study

The pDNA/APTMS-ZnO/c-GNF/ITO bioelectrode was found to be selective towards different target DNA sequences such as non-complementary, one base mismatch and complementary sequences. The selectivity was studied using EIS technique (Fig. S9). There was only slight increase in the Rct value after incubating the bioelectrode with non-complementary DNA sequence (844 Ω) (Fig. S9 curve (ii)) as compared to pDNA (curve (i)). This is very obvious since the non-complementary DNA base sequences do not match with the pDNA base sequence and no hybridization would lead to a negligible change in the charge distribution of electrode surface and therefore very less change in Rct value. When the pDNA/APTMS-ZnO/c-GNF/ITO bioelectrode was incubated with one-base mismatch DNA sequence (curve iii), there was a slight increase in the Rct value in comparison to pDNA/APTMS-ZnO/c-GNF/ITO bioelectrode which can be clearly seen by the semicircular diameter of the curve (i) and curve (iii) of Fig. S9. This might be due to partial hybridization of pDNA with bases of one-base mismatch DNA as a result of which there is increase in negative charge at the electrode surface. The negatively charged electrode surface now applied a blocking effect i.e. repulsion towards the redox ions. While on exposure of pDNA/APTMS-ZnO/c-GNF/ITO bioelectrode with cDNA there is complete hybridization of bases takes place leading to increase in negative charge and higher blocking effect towards the redox ions (curve iv) as can be ascertained from high Rct value.

Specificity study

The specificity of pDNA/APTMS-ZnO/c-GNF/ITO bioelectrode was checked towards the isolated DNA from the panel strains of *E. coli* and various other water-borne pathogens such as *Salmonella Typhimurium*, *Neisseria*

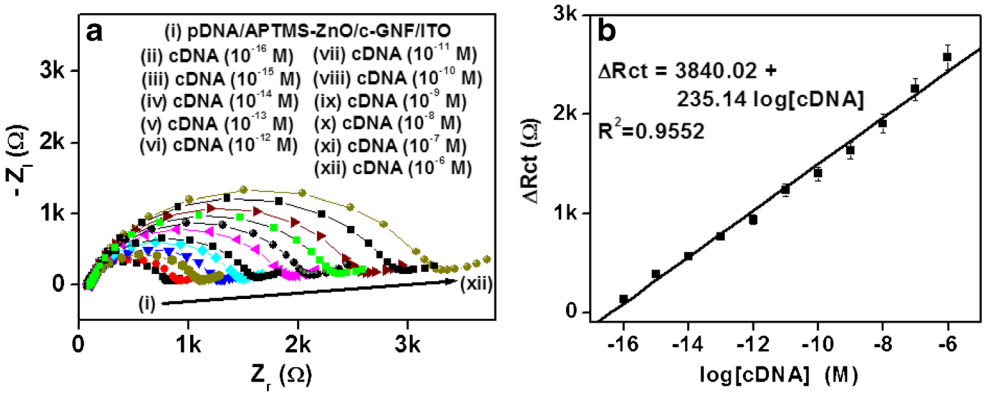


Fig. 3 **(a)** EIS data for detection of cDNA hybridized on pDNA/APTMS-ZnO/c-GNF/ITO bioelectrode, **(b)** the calibration plot for detection of cDNA on pDNA/APTMS-ZnO/c-GNF/ITO bioelectrode (in 5 mM $[Fe(CN)_6]^{3-/4-}$, PBS, 100 mM, pH 7.0, 0.9% NaCl)

meningitides and *Klebsiella pneumonia* using EIS study (Fig. S10). There was a marked increase in Rct value when the bioelectrode was incubated with clinical *E. coli* strains indicating the complete hybridization. Whereas, negligible change was recorded when the electrodewas incubated with other pathogenic stains *S. Typhimurium*, *N. meningitides* and *K. pneumonia.* While the strains of other bacteria such as *S. Typhimurium* do not affect the hybridization of the bioelectrode with the strains of *E. coli* revealing the specificity of the fabricated biosensor.

The reusability of the fabricated electrode was evaluated by immersing the electrode in a buffer solution (pH 8.0) containing Tris buffer (10 mM) and EDTA (1 mM) at 100 °C for 5 min, followed by cooling in an ice bath for about 25 min. This process completely removes the cDNA via thermal denaturation. The Rct value of the nucleic acid sensor was found to decrease after each regeneration process (5 cycles) with a total loss of 10.42% of the original Rct value. This reduction in signal may be due to surface fouling during the regeneration process indicating that the biosensor can be used for atleast 5 times (data not shown).

Conclusion

In summary, the work emphasizes the synthesis of APTMS functionalized ZnO-carboxylated graphene nanoflakes composite using wet chemical route and its electrophoretic deposition on ITO supported electrode at 120 V. This fabricated platform was immobilized with probe DNA sequences for the selective determination of *E. coli* O157:H7 DNA using impedemetric technique. The proposed *E. coli* O157:H7 DNA sensor retained its activity for 5 repetitive uses with a loss of approximately 10% activity. The lowest concentration of *E. coli* O157:H7 DNA detection using the fabricated biosensor was found to be 0.1 fM. The sensing platform can be interestingly used for the detection of other water borne pathogenic DNA sequences. Besides all this, efforts are being made to develop affinity based biosensors (using antibody–antigen interactions) to allow the detection of *E. coli* O157:H7 cells directly from water samples.

Acknowledgements N.J. and S.S. are thankful to CSIR, India for the award of SRF and RA, respectively. C.M.P. acknowledges the Department of Science and Technology, New Delhi, India for DST-INSPIRE Faculty Award (DST/INSPIRE/04/2015/000932). We thank Dr. Gajjala Sumana (CSIR-National Physical Laboratory) for interesting discussions. N.J. and I.T. are sincerely thankful to Prof. O.N Srivastava (Dept. of Physics, I.Sc. B.H.U) and Prof. Rajiv Prakash (IIT B.H.U) for providing TEM and SEM facilities respectively. B.D.M. thanks Science & Engineering Research Board (Govt. of India) for the award of a Distinguished Fellowship (SB/S9/YSCP/SERB-DF/2018).

Compliance with ethical standards

Conflict of interest The authors declare no conflict of interest.

References

1. Pandey CM, Dewan S, Chawla S, Yadav BK, Sumana G, Malhotra BD (2016) Controlled deposition of functionalized silica-coated zinc oxide nano-assemblies at the air/water interface for blood cancer detection. Anal Chim Acta 937:29–38
2. Zhang P, Wang S (2014) Designing fractal nanostructured biointerfaces for biomedical applications. Chem phys chem 15: 1550–1561
3. Xu Q, Wang L, Lei J, Deng S, Ju H (2013) Platinum nanodendrite functionalized graphene nanosheets as a non-enzymatic label for electrochemical immunosensing. J Mater Chem B 1:5347–5352
4. Wang J (2005) Nanomaterial-based electrochemical biosensors. Analyst 130:421–426
5. Pan C, Zhu J (2009) The syntheses, properties and applications of Si, ZnO, metal, and heterojunction nanowires. J Mater Chem 19: 869–884
6. Rahman MM, Ahammad AJS, Jin JH, Ahn SJ, Lee JJ (2010) A comprehensive review of glucose biosensors based on nanostructured metal-oxides. Sensors 10:4855–4886
7. Solanki PR, Kaushik A, Agrawal VV, Malhotra BD (2011) nanostructured metal oxide-based biosensors. NPG Asia Mater 3:17–24 (2011)
8. Chai G, Lupan O, Chow L, Heinrich H (2009) Crossed zinc oxide nanorods for ultraviolet radiation detection. Sens Actuators A: Physical 150:184–187

9. Mehrabian M, Azimirad R, Mirabbaszadeh K, Afarideh H, Davoudian M (2011) UV detecting properties of hydrothermally synthesized ZnO nanorods. Physica E Low dimens Syst Nanostruct 43:1141–1145
10. Ali SMU, Nur O, Willander M, Danielsson B (2010) A fast and sensitive potentiometric glucose microsensor based on glucose oxidase coated ZnO nanowires grown on a thin silver wire. Sensors Actuators B Chem 145:869–874
11. Fulati A, Ali SMU, Asif MH, Alvi NH, Willander M, Brännmark C, Stralfors P, Börjesson SI, Elinder F, Danielssond B (2010) An intracellular glucose biosensor based on nanoflake ZnO. Sensors Actuators B Chem 150:673–680
12. Mohammed AM, Ibraheem IJ, Obaid AS, Bououdina M (2017) Nanostructured ZnO-based biosensor: DNA immobilization and hybridization. Sens Biosensing Res 15:46–52
13. Ahmad R, Tripathy N, Jang NK, Khang G, Hahn YB (2015) Fabrication of highly sensitive uric acid biosensor based on directly grown ZnO nanosheets on electrode surface. Sens. Actuators B Chem 206:146–151
14. Tak M, Gupta V, Tomar M (2014) Flower-like ZnO nanostructure based electrochemical DNA biosensor for bacterial meningitis detection. Biosens Bioelectron 59:200–207
15. Chen X, Yan H, Shi Z, Feng Y, Li J, Lin Q, Wang X, Sun W (2017) A novel biosensor based on electro-co-deposition of sodium alginate-Fe3O4-graphene composite on the carbon ionic liquid electrode for the direct electrochemistry and electrocatalysis of myoglobin. Polym Bull 74:75–90
16. Salih E, Mekawy M, Hassan RYA, El-Sherbiny IM (2016) Synthesis, characterization and electrochemical-sensor applications of zinc oxide/graphene oxide nanocomposite. J Nanostructure Chem 6:137–144
17. Tiwari JN, Tiwari RN, Kim KS (2012) Zero-dimensional, one-dimensional, two-dimensional and three-dimensional nanostructured materials for advanced electrochemical energy devices. Prog Mater Sci 57:724–803
18. Pokropivny VV, Skorokhod VV (2007) Classification of nanostructures by dimensionality and concept of surface forms engineering in nanomaterial science. Mater Sci Eng C 27:990–993
19. Nel AE, Mädler L, Velegol D, Xia T, Hoek EMV, Somasundaran P, Klaessig F, Castranova V, Thompson M (2009) Understanding biophysicochemical interactions at the nano–bio interface. Nat Mater 8:543–557
20. Loo AH, Bonanni A, Pumera M (2013) Thrombin aptasensing with inherently electroactive graphene oxide nanoplatelets as labels. Nanoscale 5:4758–4762
21. Chang H, Wang Y, Li J (2011) Electrochemical DNA sensors: from nanoconstruction to biosensing. Curr Org Chem 15:506–517
22. Chen A, Chatterjee S (2013) Nanomaterials based electrochemical sensors for biomedical applications. Chem Soc Rev 42:5425–5438
23. Scallan E, Hoekstra RM, Angulo FJ, Tauxe RV, Widdowson MA, Roy SL, Jones JL, Griffin PM (2011) Foodborne illness acquired in the United States-major pathogens. Emerg Infect Dis 17:7–15
24. Tiwari I, Singh M, Pandey CM, Sumana G (2015) Electrochemical detection of a pathogenic *Escherichia coli* specific DNA sequence based on a graphene oxide–chitosan composite decorated with nickel ferrite nanoparticles. RSC Adv 5:67115–67124
25. Jaiswal N, Pandey CM, Soni A, Tiwari I, Rosillo-Lopez M, Salzmann CG, Malhotra BD, Sumana G (2018) Electrochemical genosensor based on carboxylated graphene for detection of waterborne pathogen. Sensors Actuators B Chem 275:312–321
26. Xu M, Wang R, Li Y (2017) Electrochemical biosensors for rapid detection of Escherichia coli O157:H7. Talanta 162:511–522
27. Wang J, Li S, Zhang Y (2010) A sensitive DNA biosensor fabricated from gold nanoparticles, carbon nanotubes, and zinc oxide nanowires on a glassy carbon electrode. Electrochim Acta 55: 4436–4440
28. Solanki PR, Kaushik A, Chavhan PM, Maheshwari SN, Malhotra BD (2009) Nanostructured zirconium oxide based genosensor for *Escherichia coli* detection. Electrochem Commun 11:2272–2277
29. Yang Y, Wang Z, Yang M, Li J, Zheng F, Shen G, Yu R (2007) Electrical detection of deoxyribonucleic acid hybridization based on carbon-nanotubes/nano zirconium dioxide/chitosan-modified electrodes. Anal Chim Acta 584:268–274
30. Rosillo-Lopez M, Lee TJ, Bella M, Harta M, Salzmann CG (2015) Formation and chemistry of carboxylic anhydrides at the graphene edge. RSC Adv 5:104198–104202
31. Wu C, Qiao X, Chen J, Wang H, Tan F, Li S (2008) A novel chemical route to prepare ZnO nanoparticles. Mater Lett 60: 1828–1832
32. Zhu J, He J (2012) Facile synthesis of Graphene-wrapped honeycomb MnO_2 Nanospheres and their application in Supercapacitors. Appl Mater Interfaces 4:1770–1776
33. Ivers-Tiffée E, Weber A, Schichlein H (2010) Electrochemical impedance spectroscopy. Handbook of Fuel Cells, John Wiley Sons Ltd Chichester
34. Khan SB, Faisal M, Rahman MM, Jamal A (2011) Low-temperature growth of ZnO nanoparticles: Photocatalyst and acetone sensor. Talanta 85:943–949

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Wireless Personal Communications
https://doi.org/10.1007/s11277-019-07005-6

An Improved Authentication Protocol for Wireless Body Sensor Networks Applied in Healthcare Applications

Kakali Chatterjee[1]

© Springer Science+Business Media, LLC, part of Springer Nature 2019

Abstract
Wireless body sensor networks are now very popular for modern healthcare applications like health monitoring, remote healthcare, emergency healthcare etc. This healthcare application mainly used for constant monitoring of health specific data while doing regular activities. The wireless body area network is consisted of small wearing sensors which are implanted in human body for capturing medical data and send to a medical server through a network connector to base station. A major issue is secure transmission of patient's medical healthcare information to the medical server through wireless communication. Also the data collected in this network is very sensitive because on the basis of this clinical data, further treatment will be occurred. Therefore, security requirements such as confidentiality, integrity, authenticity should be guaranteed during communication. This paper proposes a strong mutual authentication protocol based on public key cryptography for satisfying all security requirements. The proposed authentication approach resists the major vulnerable attacks in wireless body sensor networks with low computational and communicational load.

Keywords Mutual authentication · Paillier cryptosystem · Body sensor networks

1 Introduction

Recent days healthcare applications are playing a major role for remote health monitoring. These applications are mainly used the patient data which is captured by the wearable sensors. These wearable sensors are the part of a Wireless Body Area Networks for patient monitoring purpose. The implanted sensors will collect the data from the patient body and the smart phone will collect the data from the sensors to supply it to the medical server [1].

Originally the wireless body area network (WBAN) was proposed by Zimmerman [2] uses wireless personal area network (WPAN) technology. In current days, the WBAN has played an important role in research community and health organizations for smart health applications. Normally in this type of networks many low power intelligent sensors are placed in or around the human body for measuring vital parameters such as heart

✉ Kakali Chatterjee
kakali2008@gmail.com

[1] CSE Department, NIT Patna, Patna, India

Published online: 12 December 2019

rate, pulse, blood pressure, and temperature and oxygen saturation etc. Real-time monitoring could be implemented remotely through these sensors. A body sensor based healthcare application can provide many benefits in health sector also. This type of application can provide a convenient environment for monitoring those vital parameters during daily lives activity and medical conditions for long term critically ill (paralyzed patient, cancer patient) person to reduce the huge hospitalization cost. The sensed data is stored in health cloud so that doctors can view these data at anytime from anywhere in the country. In such environment, one of the biggest challenges is to send the data securely to the cloud server because any changes of this sensitive data may cause wrong diagnosis and wrong treatment of the patient. As the data gathered from the sensors will send to the base stations through a personal server (network coordinator), an adversary can easily capture the data from the wireless channels and modify that. This wrong data could endanger patient's life. Also the personal server can be captured by the adversary with wrong intension. Hence strong authentication and secured communication is essential in this scenario. This is the main objective of this research work.

In WBAN, several authentication schemes based on Receiver Signal Strength [3–6] and Proximity [7–9] and Biometric based [10–12] has been found in the literature. The drawback of these schemes is mainly depending upon the distance between devices (sensors) which must be half of the wavelength distance of each other. These drawbacks can be overcome using cryptography based scheme because those schemas mainly depend upon the operational complexity not with signal strength or location. But implementation of suitable cryptographic protocol in this environment is challenging as these sensor devices are constrained in terms of memory, computing power and energy supply.

The traditional public key cryptography based authentication schemes are also found in literature [13–17]. One of the most secure public key cryptographic techniques is Elliptic Curve Cryptography which provides the highest security with smaller key size [14]. Compared to traditional techniques, ECC is more suitable for sensor networks due to the constraints like limited computing capabilities and battery capacity. In general practical implementations, ECC public key is distributed through digital certificates with the help of public key infrastructure (PKI). The distribution and management of certificates increases communicational overload. Thus ECC based authentication schemes [18–22] are not suitable for their complex operations. Shamir [23] first proposed Identity based cryptography where the identity plays the role of the public key. Therefore, the ID-based cryptography mainly overcomes the certificates management problem of the trusted third party. Yang et al. [24] proposed an ID based mutual authentication scheme for mobile devices. Yoon et al. [25] shows the demerits of this scheme through cryptanalysis. He et al. [26] used ECC to design a new ID based authentication scheme which is provably secure. Later on Wang et al. [27] pointed that He et al.'s scheme suffers parallel session attack, reflection attack and also does not provide mutual authentication. Biswas et al. [28] proposed another ECC based authentication to remove security weakness in Yoon and Yoo's scheme. But Truong et al. [29] found that Yoon et al. scheme could not resist the denial of service attack. Some light weight authentication scheme have been found in wireless sensor networks [30–32] which fails to maintain perfect forward secrecy in communication.

Although the above ID-based authentication schemes is mainly suitable for client server environment with better performance than earlier schemes, but not suitable for BSN due to the algorithmic complexity. This paper proposes an improved authentication scheme based on public key cryptography for not only data security but also to maintain the privacy in BSN. The proposed scheme is based on Paillier cryptosystem which have homomorphic and self-binding property. This keeps the sensed data private while third party processes

the data without seeing it which is essential for body sensor network. Also this cryptosystem provides randomness in encrypted data, so that the same plain text after several encryptions produces different cipher texts. Hence the main contribution in this paper is as follows:

- The proposed scheme is efficiently authenticates user device using biometric features while data transmission from body sensors to health server in WBAN.
- This scheme used Pallier cryptosystem which is mainly used for privacy preservation of sensor data.
- Proposed Authentication approach enforces very light computational load due to simple operations like exclusive-OR operations.
- Mutual authentication as well as secret session key generation is essentially improves security.
- Proposed scheme also has low communicational load and messages are encrypted using symmetric encryption for this critically constrained devices.

The rest of the paper is organized as follows:
Section 2 presents Background, Sect. 3 provides Proposed Mutual Authentication Protocol; Sect. 4 discusses Security Analysis; Sect. 5 presents Performance Analysis; Finally, concluded in Sect. 6.

2 Background

In this section security threats and security requirements in WBAN are discussed.

2.1 Security Requirements, Threats and Solutions in WBAN

There are many sensor nodes in WBSN which are involve in data sensing. After that the gateway node will collect the sensed data and transmitted it to a health server via internet [33]. In such scenario, following security requirements are found in WBSN:

- Data Confidentiality is essential to protect the data from a disclosure, the system require data confidentiality.
- Data integrity is necessary as an adversary can alter the data that is transmitted over an insecure channel.
- Data authentication is essential to ensure that the data is coming from trusted node.
- Data availability is to ensure that the patient's information is accessible to the doctor.
- Efficient key management is essential to securely transmission of data
- Efficient encrypting techniques are essential as the computation power is limited.
- Anonymity and Non traceability is required to maintain the privacy of the patient data.
- Perfect Forward Secrecy must be maintained even if the secret keys of the client and application provider is compromised.
- Mutual Authentication is essential to ensure user and server.

WBANs are vulnerable to many threats such are listed below:

- Message Modification- This type of threat is applicable to the message after intercepting it from the network. Sometimes the adversary can delete or delay the data to harm the person. Strong encryption algorithm and hash function can be used for prevention of this attack.
- Message Disclosure- This threat is activated when an application fails to properly protect sensitive information from others. Due to this privacy threat, some sensitive private data may be disclosed to those who are not supposed to get the information.
- Unauthorized access- When some unauthorized user gets some access grant, the data and the resources can be misused. Unauthorized access can be prevented by strong access control policy.
- Denial of Service- Due to this attack, the user cannot get the desired service. The attacker might use different strategies to achieve this such as by inserting bogus request to a server or drop request packets of the client. It can be resisted by implementing suitable Intrusion Detection System.
- Node tampering- An adversary can gain full control over a sensor node by this attack.

Also he can perform malicious activity by capturing that node. It can be resisted by any inconsistency detection algorithm.

- Routing attacks- Fake routing messages can be inserted in communication channels by this attack to divert the routes of the messages. Hence it is preferable to use secure routing protocols.
- Jamming Attacks- Due to this attack, the adversary can perform radio frequency interference to block the entire network. One simple solution is to apply high transmission power on jammed channels to avoid this attack.

2.2 Paillier Cryptosystem

The Paillier Cryptosystem is a modular, public key encryption scheme created by Pascal Paillier, with several interesting properties [34]. This public key encryption is based on composite residuosity classes. It has homomorphic property through which a person can delegate to process his own data without giving access on it. Also, it has self-blinding property through which a plaintext can be mapped into many different ciphertexts. Like all cryptosystems, it has three processes: Key generation, encryption and decryption.

In the key generation process, to construct the public key, one must choose two large primes, p and q, then calculate their product, n = p.q. Then a semi-random, nonzero integer, g, in Z_n^2, must be selected, such that the order of g is a multiple of n in Z_n^{*2} [Z_n^{*2} being the units, or invertible elements, of Z_n^2].

The modular paillier cryptosystem uses the concept of public key encryption technique. To encrypt the message, the sender first generates a public key by choosing two large primes number p and q. Now, calculates n = p*q. For example [34], consider p = 7 and q = 11. The calculated value of n is 77. After that, a semi-random, non-zero integer g is chosen such that the order of g is a multiple of n in Z_n^{*2}. The choose value of g is 5652, which fulfill the addressed properties. Thus, the generated public key is represented by (n,g) = (77, 5652).

Steps for Encryption:

1. Generate the public key.

In this example, the generated public key is (77, 5652).

2. Create a message m such that $m \in Z_n$.
   Let $m = 42$.
3. Choose a random nonzero integer $r \in Z_n$.
   Let $r = 23$.
4. Compute the ciphertext $c \equiv g^m r^n \bmod n^2$.
   The value of $c \equiv (5652)^{42} * (23)^{77} \bmod 5929 \equiv 4624 \bmod 5929 = 4624$

Steps for Decryption:

1. Compute $\lambda(n) = \text{lcm}[(p-1)(q-1)]$ by Carmichael's function.
   $\lambda(77) = \text{lcm}[(6),(10)] = 30$
2. Calculation of $u = g^{\lambda(n)} \bmod n^2$ is necessary for the decryption of the message.
   $U = (5652)^{30} \bmod 5929 = 3928$
3. $L(u) = (u-1)/n$
   $L(3928) = 3927/77 = 51$
4. $\mu = L(u)^{-1} \bmod n$
   $\mu = 51^{-1} \bmod 77 = 74$
5. Compute the plain text message $m \equiv L(c^{\lambda(n)} \bmod n^2) * \mu \bmod n$
   $M \equiv L(4624^{30} \bmod 5929) * 74 \bmod 77 \equiv 42 \bmod 77 \equiv 42$ (plain text message)

The paillier cryptosystem have following properties:

1. The result of multiplication of two encrypted is equivalent to the addition of original of plaintext mod n.
2. The result of full encryption of the second message (r^n is left out) is multiply with the first encrypted message is equivalent to the addition of original of plaintext mod n.
3. A constant power on the ciphertext gives the result in the term of constant multiple of the original plaintext.

To solve Privacy issues, Paillier cryptosystem's are used to ensure homomorphic encryption in databases [35]. Pallier cryptosystems are mainly depending upon modular arithmetic operations and integer arithmetic operations such as multiplication and exponentiation modulo n^2 for the encryption and decryption. Details of module sizes of Pallier Cryptosystem are as follows (Table 1):

Table 1 Fundamental sizes of Pallier cryptosystem

Pallier operations	Modulo size n	Modulo size n^2	# of ModMult with n^2	# of words [d] for n^2
Encryption/Decryption	512	1024	516/514	32
Encryption/Decryption	1024	2048	1028/1026	64
Encryption/Decryption	2048	4096	2052/2050	128

3 Proposed Authentication Protocol

In this section an authentication protocol based on Paillier Cryptosystem has been proposed.

3.1 System Network Model

This is a system network model of an electronic health care system which is mainly monitoring vital parameters of critically ill patient. The doctors are monitoring this parameter remotely with the help of their smart phones. Considering this scenario, many body sensor nodes has been installed in the patient body for measuring different vital parameters like heart rate, pulse, pressure, temperature and oxygen saturation etc. All sensor nodes are sending captured data to the gateway node. The gateway node will send data to local processing centre (Base Station) through the network coordinator which is called User Device. The user device is issued by the health care system while initial phase. All the captured data is gathered in the health cloud through this device and the doctors, nurses are observing data from the cloud server. The main focus is only on the secure transmission of data from the body sensor to the health cloud server. Figure 1 shows the architecture of the network.

3.2 Notations Used in this Protocol

Consider two large prime p and q over ($\mathbb{F}_p$).

N	The product of two prime p and q
G	The semi-random, nonzero integer in $\boldsymbol{Z}_{\mathbf{n}}^{2}$
H(.)	Hash function which is a one way function
$E_k(.)$	Asymmetric encryption function which uses key K
$D_k(.)$	Asymmetric decryption function which uses key K
T_1, T_2	Present Timestamp of node

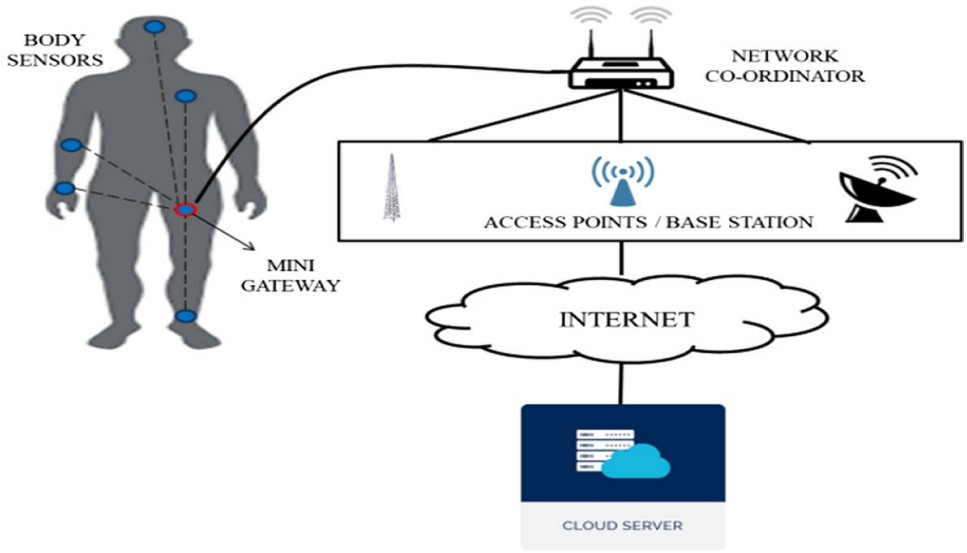


Fig. 1 System model of BSN

ΔT	Expected transmission delay
T_s	Session Duration Time
$\oplus$	Logic exclusive OR (XOR) operation
$\|$	A bitwise concatenation operation
UID	Identity of User Device
SID	Health Server ID
PID	Patient ID
P	Secret No
n_u, P_u	Private and Public key of User Device respectively
n_s, P_s	Private and Public key of Health Server respectively

Assumptions

Consider many body sensor node has been installed in a human body for measuring different body parameters. All nodes are sending captured data to the gateway node. The gateway node will send data to local processing center (Base Station) through the network coordinator which is called User Device. After deployment of the sensor nodes, following assumptions are made

- After valid registration of the user device, it will forward capture data of sensor node.
- User Device is considered as trusted one.
- Clock of the nodes in human body are synchronized with the user device.
- Each sensor data is accumulated to a main gateway which is also a sensor.
- After Registration BS will send user details to Health Server.
- Health Server will create each patients table for storing his/her data.
- BS is playing the role of trusted third party.
- Signal intensity is within a given fixed range with high transmission power.

3.3 Phases of Authentication Protocol

The proposed scheme shown in Fig. 2 is divided into three phases as described below:

A. User Registration phase:

First the patients biometric input B will capture and pass through a fuzzy extractor which could produce a random string σ. The Fuzzy extractor will operate two functions (*Gen & Rep*) which will produce random string for identification. The random string will be same for close change of input. Actually the *Gen()* will output a random string σ and a random auxiliary string ν in enrolment phase. *Rep()* is computed while identification process after receiving biometric input B^* and the corresponding auxiliary string ν to recover σ.

(i) During registration phase, a temporal key is generated from the hash of biometric profile. The device first capture patients biometric impression B_i and computes $(\sigma_i,\nu_i) = Gen(B_i)$. The temporal key is $h(PW \| \sigma_i)$ where PW is the user password.

Now user device will send a request which contain user identity (UID) and the temporal key (η) to base station (Local Processing Centre).

The request message $R_u = UID \| h(PW \| \sigma_i)$

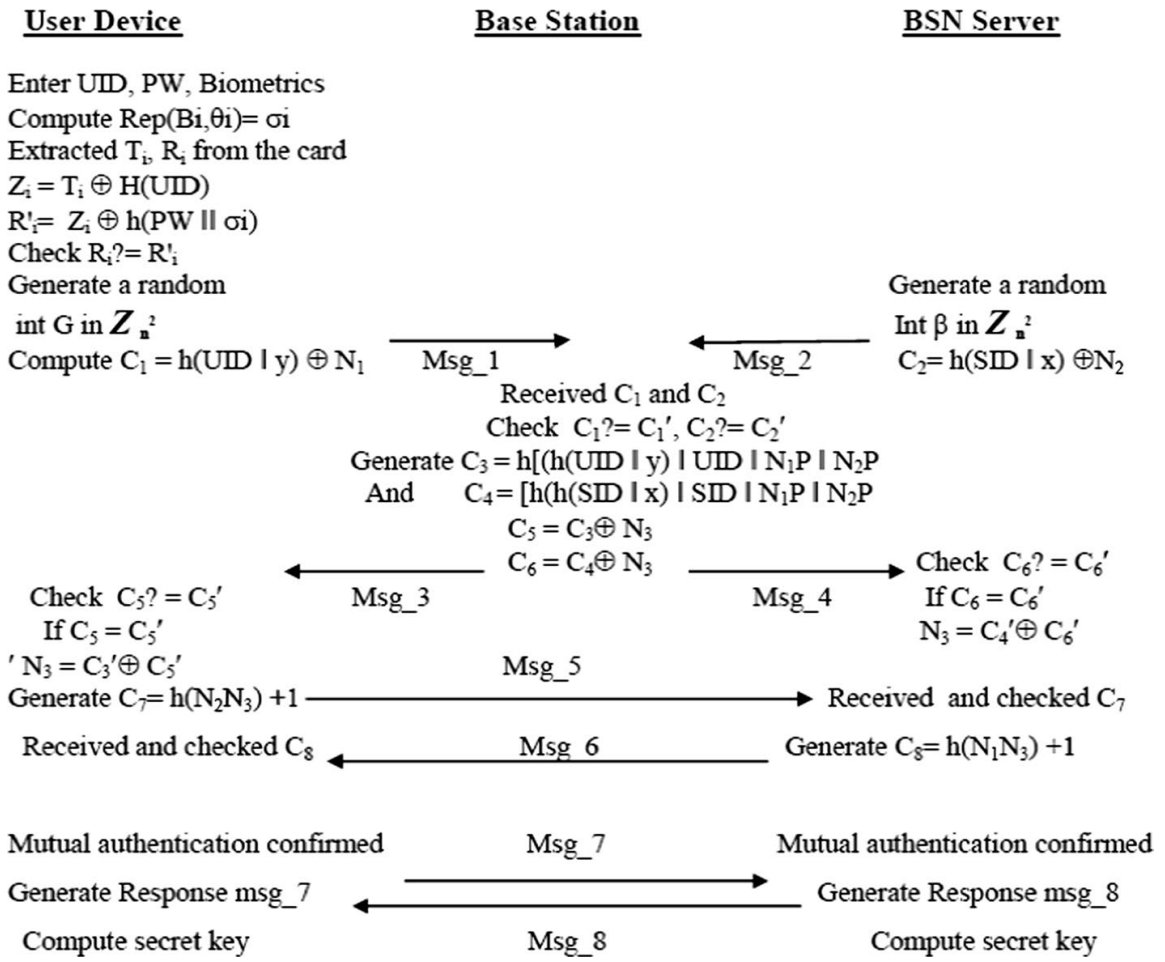


Fig. 2 Proposed authentication scheme

At the same time BSN cloud server send a registration request $R_s \| SID$ where $R_s = h(SID \| x)$. Here x is the server secret.

(ii) After R_u and R_s local processing center will compute the following parameters

$$R_i = h(UID \| y)$$

$$Z_i = R_i \oplus h(PW_i \| \sigma_i)$$

$$T_i = Z_i \oplus h(UID)$$

T_i and R_i is stored in the Smart card and the card is issued for that user device which will collect patient data to the local processing center.

(iii) Now the base station also sends a list of registered users to cloud health server which is responsible to collect and store data of the patients. In this way the base station completes the registration phase for secure data collection. The message flow is shown in Table 2. BS will send an updated list of records to cloud server including UID and PID (Patient ID) with regular interval.

B. Login phase:

During this phase, user will perform following steps as shown in Table 3:

Table 2 Flow of registration phase

Steps	Flow	Message	Description
1.	UN → BS	Sub_id()	UID ‖ h(PW ‖ σ_i)
2.	HS → BS	Reg_no()	h(SID‖ x)
3.	BS → HS	Nod_list()	UID ‖ PID ‖P
4.	BS → UN	Con_no()	UID ‖ P ‖ R_s

Table 3 Flow of Login phase

Steps	Flow	Message	Description
1.	UN → BS	Msg_1	UID‖ C_1‖ PU_A‖ N_1P
2.	HS → BS	Msg_2	SID ‖ C_2 ‖ PU_S‖ N_2P
3.	BS → UN	Msg_3	Enc(C_3 ‖ C_5 ‖ N_2P)
4.	BS → HS	Msg_4	Enc(C_4 ‖ C_6 ‖ N_1P)

Step 1 The user device first submits its biometric impression B_i^* and UID with password (PW). The device will compute $Gen\,(B_i^*, v_i) = \sigma_i$ and extracted the value T_i from the card to compute $Z_i' = T_i \oplus h(UID)$.

Now the device computes $R_i' = Z_i' \oplus h\,(PW_i||\sigma_i)$ and matches with extracted value of R_i.

Step 2 The user device will generate the public key $PU_A = (N,G)$ using Pailliar cryptosystem. Now it sends a message to local processing center (BS) which contains a token $C_1 = h(UID||x) \oplus N_1$ and the public key with user ID for authentication.

Hence Msg_1 $= UID\|C_1\|PU_A\|N_1P$

Similarly Health server will send a message Msg_2 $= SID\|C_2\|PU_S\|N_2P$

Step 3 BS will verify $C_1? = C_1', C_2? = C_2'$. After verification BS will generate

Two tokens for user device and health server. The tokens are

$$C_3 = h\left[h(UID||y)\right]||UID||N_1P||N_2P \text{ and } C_5 = C_3 \oplus N_3$$

$$C_4 = h[h(SID||x)]||SID||N_1P||N_2P \text{ and } C_6 = C_4 \oplus N_3$$

BS will generate two encrypted Msg using Pailliar cryptosystem for user device and server.

$$Msg_3 = Enc\,(C_3||C_5||N_2P)$$

$$Msg_4 = Enc\,(C_4||C_6||N_1P)$$

Step 4 User Device first checks $C_5? = C_5'$. If it is equal than extract N_3 by $C_3 \oplus C_5$.

Now generate another token $C_7 = h(N_2N_3) + 1$ for acknowledgement.

Similarly Server first checks $C_6? = C_6'$. If it is equal than extract N_3 by $C_4 \oplus C_6$.

Now generate another token $C_8 = h(N_1N_3) + 1$ for acknowledgement.

C. Mutual authentication phase:

In this phase the user device and BSN server will perform mutual authentication by using following steps:

Step 4 In this step User Device generate a message Msg_5 = Enc [UID || C_7|| T_1] and send to BSN server.

After receiving of Msg_5, server decrypts it and get the session time. If $T - T_1 \leq \Delta T$ where T is the present timestamp of server, then it accepts the message otherwise rejects. Now BSN servers get the UID value and compute

$$C_7' = h(N_2N_3) + 1 \text{ . If } C_7' = C_7, \text{ then it computes Msg_6} = \text{Enc}\,[\text{SID}||C_8||T_2]$$

Step 2 In this step, User received Msg_6 from BSN server. After decryption it first checks the timestamp with present timestamp. If it is valid, then it computes $C_8' = h(N_1N_3) + 1$.

If $C_8' = C_8$, then it generates the response message Msg_7 as a proof of mutual authentication and shared secret key generation.

$$\text{Msg_7} = h(C_7)||\text{UID and the shared secret key } K = h(\text{UID}||y) \oplus h(\text{SID}||x) \oplus h(N_3).$$

Step 3 After receiving Msg_7, server confirmed that the user is valid user and generate the shared secret key K = h(UID|| y) ⊕ h(SID|| x) ⊕ h(N_3). Server sends the confirmation message Msg_8 = h(C_8)|| SID. The message flow is shown in Table 4. After that all messages will be sending using symmetric encryption.

4 Detail Security Analysis

In this section analysis of the proposed authentication scheme on the basis of different attacks and also formal proof using BAN logic is performed.

4.1 Probable Attack Analysis

This sub section includes detail discussion of following attacks:

4.1.1 Password Guessing Attack

Considering a case, an attacker gets the smart card and he has tried to login the system. First he need to give the biometric impression of the patient as the device is registered for that patient. The smart card also carries the specific biometric parameter of that patient. If he is able to do that, then also he has to submit the password for authentication. Moreover

Table 4 Flow of mutual authentication phase

Steps	Flow	Message	Description
1.	UN → CS	Msg_5	Enc [UID || C_7|| T_1]
2.	CS → UN	Msg_6	Enc [SID || C_8|| T_2]
3.	UN → CS	Msg_7	h(C_7)|| UID
4.	CS → UN	Msg_8	h(C_8)|| SID

it is not possible to guess biometrics and password correctly at the same time. Hence online password guessing attack could not be successful.

The off-line password guessing attack also will not work as $h(PW \| \sigma_i)$ is used to calculate R_u and the two parameters

$$Z_i = R_i \oplus h(PW_i||\sigma_i) \text{ where } R_i = h(UID\|y),\ T_i = Z_i \oplus h(UID)$$

At a time both are required and limited attempts will be given. Hence this protocol will resist dictionary attack.

4.1.2 Replay Attack

To prevent replay attack, nonce value is inserted in each message. Also the freshness of the message is always be calculated after receiving each message. For example $Msg_1 = UID\| C_1\| PU_A\| N_1P$ contains random nonce for preventing replay attack.

Similarly Health server message $Msg_2 = SID \| C_2 \| PU_S\| N_2P$ also carrying random nonce for preventing replay attack.

In mutual authentication phase, User Device generate a message $Msg_5 = Enc\ [UID \| C_7\| T_1]$ and send to BSN server. After receiving of Msg_5, server decrypts it and get the session time. If $T–T_1 \leq \Delta T$ where T is the present timestamp of server, then it accepts the message. This is also for prevention of replay attack.

4.1.3 Impersonation Attack

Impersonation attack happened when an attacker pretend as a valid user. The probability of this attack is very less as biometric impression is captured and inserted in the smart card. To prevent impersonation, mutual authentication phase is also there. In this phase the BSN server authenticates itself to the user so that any fake server cannot able to establish a connection.

4.1.4 Insertion/Message Modification Attack

In this authentication protocol after login message all messages are in encrypted form. After mutual authentication, a shared session key is generated to encrypt the data which the sensors are sending to BSN server. For example,

$$Msg_3 = Enc\,(C_3||C_5||N_2P)$$

$$Msg_4 = Enc\,(C_4||C_6||N_1P)$$

$$Msg_5 = Enc\,[UID\|C_7||T_1]$$

$$Msg_6 = Enc\,[UID\|C_8||T_2]$$

Now after mutual authentication the shared secret key is used for data transmission. For insertion attack, an attacker must have to decrypt the messages and for decryption he needs the shared key. If any insider knows the UID and SID, then also he has to know N_3. No single message carrying N_3 which can be tracked by an attacker. Hence this attack cannot be possible.

4.1.5 Man-in-Middle Attack

In this attack, the attacker manages to set a key between the user and the server so that he will be able to hear all the transmitted messages. For that intension, he wants to set a common key between user device and BSN server. Now to replace the value of attacker replace the value of C_7, C_8 by C_7', C_8', the attacker first have to decrypt the Msg_5 and Msg_6. It is impossible. If he replaces these messages with another encrypted messages, then also a successful session key will not established.

Thus this attack is not possible with this scheme.

4.1.6 Server Spoofing Attack

There is no verification table stored in the server so that it can authenticate any user device. All user devices are authenticated by the Base Station and after authentication this device will communicate with the server. Suppose the intended server block the BSN server and capture the message $\text{Msg_4} = \text{Enc}(C_4 \parallel C_6 \parallel N_1 P)$ which is coming from the user device. The attacker must decrypt the message to get C_4 and C_6 which carries N_3. Again to authenticate itself, the fake server also needs to know the value of x as the authentication needs h(SID||x). Therefore, Server Spoofing attack cannot be successful.

4.1.7 Perfect Forward Secrecy

This property shows that even if the secrets x, y of the past session is disclosed, then also the attacker cannot able to calculate the past session key. The shared secret key $K = h(\text{UID} \parallel y) \oplus h(\text{SID} \parallel x) \oplus h(N_3)$. To calculate the past session key, the attacker must know the past nonce value of N_3. Hence the scheme is preserving perfect forward secrecy.

4.1.8 Message Disclosure

This threat is triggered when some sensitive private data may be disclosed to those who are not supposed to get the information. During communication, one of the sensitive information is login details. In this scheme, password $h(PW \parallel \sigma_i)$ is used to calculate R_u and the two parameters $Z_i = R_i \oplus h(PWi \parallel \sigma_i)$where $R_i = h(UID\|y)$, $T_i = Z_i \oplus h(UID)$. From this message it is very difficult to gain the knowledge of password which is sensitive information.

Other sensitive information such as patient vital parameters data is always sent and restored in an encrypted form so that unauthorized person cannot get the data. Only those who have the shared secret key and access right, can access the original data. Pailliar cryptosystem is used to maintain the privacy of the data. Hence message disclosure threat cannot be successful in the proposed scheme.

4.1.9 Node Tempering Attack

In this proposed scheme, the sensor nodes are implanted inside the body. Suppose the user device which is transferring the data to the base station is compromised and the attacker gets all stored information such as UID, PW. From that, T_i, R_i is generated and stored in the Smart card and the card is issued for that user device.

During login, the user device first submits user's biometric impression B_i^* and UID with password (PW). The device will compute $Gen\ (B_i^*, \nu_i) = \sigma_i$ and extracted the value T_i from the card to compute $Z_i' = T_i \oplus h(UID)$. Now the device computes $R_i' = Z_i' \oplus h(PW_i || \sigma_i)$ and matches with extracted value of R_i.

If the device is tampered, then also with valid smart card any adversary cannot able to send data using that device.

4.1.10 Jamming Attack

It is one of the possible attacks in sensor network. To avoid this attack, high transmission power on jammed channels has been considered as mentioned in protocol assumptions.

4.2 Authentication Proof Based on BAN Logic

The authentication protocol can be proved by using BAN logic [14] which is defined as a set of logical rules to analyze any protocol. Here goal has been set and on the basis of six defined rule. For verification, this work first starts with its normal definition:

Let us consider R and S are principals, I and J are statements, K is the encryption key.

Now the standard relationships and its uses are shown in Table 5 below

Table 5 Symbol representation

$R \mid \equiv I$	R believes I
$\#(I)$	I is fresh as no one has sent it before
$R \rightarrow I$	R controls I
$R \Delta I$	R receives message with I which he can read
$R \mid \sim I$	R sent a message containing I sometime
(I, J)	I or J is one part of the formula (I, J).
$\langle I \rangle J$	The formula I combines with a secret parameter J
$\{I\}_K$	The formula I is encrypted with the key Key
$(I)_h$	The formula I is hashed
RKS	R and S use the shared key Key to communicate
SK	The session key used in the current session
$\frac{R \mid \equiv RKS, R \lhd \{I\}_k}{R \mid \equiv SI}$	Message meaning rule(if R believes that Key is shared with S, if R receives a message encrypted with key, then R trust on S)
$\frac{R \mid \equiv \#(I)}{R \mid \equiv \#(I,J)}$	Freshness conjugation rule (if I is fresh then message (I, J) is also a fresh)
$\frac{R \mid \equiv S \mid \equiv (I,J)}{R \mid \equiv S \mid \equiv (I)}$	The belief rule (If the principal R believes I and J, then the principal S believes (I, J)
$\frac{R \mid \equiv \#(I), R \mid \equiv S \mid \sim I}{R \mid \equiv S \mid \equiv I}$	The nonce-verification rule (If the principal R believes that I is fresh and the principal S sent I once then the principal R believes that S believes I)
$\frac{R \mid \equiv S \mid \rightarrow I, R \mid \equiv S \mid \equiv I}{R \mid \equiv I}$	The jurisdiction rule (If the principal R believes that S has jurisdiction right to I and S believes I, then R believes that I is true)

For correctness measurement, the key agreement protocol must achieve the following goals:

Goal 1: $U_n| \equiv U_n \overset{SK}{\leftrightarrow} S_n$

Goal 2: $U_n|\equiv S_n| \equiv U_n \overset{SK}{\leftrightarrow} S_n$

Goal 3: $S_n| \equiv U_n \overset{SK}{\leftrightarrow} S_n$

Goal 4: $S_n|\equiv U_n| \equiv U_n \overset{SK}{\leftrightarrow} S_n$

Now transforming the proposed scheme to the idealized form is as follows:

1. Message 1: $U_n \rightarrow \text{BS}$: $(UID, I)_{\text{h(UID}\|\text{k)}}$
2. Message 2: $S_n \rightarrow \text{BS}$: $(UID, I, SID, J)_{\text{h(SID}\|\text{k)}}$
3. Message 3: $\text{BS} \rightarrow U_n$: $(UID, SID, I, J, U_n \overset{J}{\leftrightarrow} S_n)_{\text{ h(UID}\|\text{ k)}}$
4. Message 4: $\text{BS} \rightarrow S_n$: $(UID, SID, I, J, U_n \overset{I}{\leftrightarrow} S_n)_{\text{ h(SID}\|\text{ k)}}$
5. Message 5: $S_n \rightarrow U_n$: $(UID, SID, I, J, U_n \overset{SK}{\leftrightarrow} S_n)_{SK}$
6. Message 6: $U_n \rightarrow S_n$: $(UID, SID, I, J, U_n \overset{SK}{\leftrightarrow} S_n)_{SK}$

Verifying this protocol using BAN logic requires some assumption. They are as follows

- A_1: $U_n| \equiv \#(I)$
- A_2: $S_n| \equiv \#(J)$
- A_3: $U_n| \equiv U_n \overset{\text{h(UID}\|\text{k)}}{\leftrightarrow} BS$
- A_4: $BS| \equiv U_n \overset{\text{h(UID}\|\text{k)}}{\leftrightarrow} BS$
- A_5: $S_n| \equiv S_n \overset{\text{h(SID}\|\text{k)}}{\leftrightarrow} BS$
- A_6:$BS| \equiv S_n \overset{\text{h(SID}\|\text{k)}}{\leftrightarrow} BS$
- A_7: $U_n| \equiv \text{BS} \Rightarrow U_n \overset{J}{\leftrightarrow} S_n$
- A_8: $S_n| \equiv \text{BS} \Rightarrow U_n \overset{I}{\leftrightarrow} S_n$
- A_9: $S_n| \equiv U_n \Rightarrow U_n \overset{SK}{\leftrightarrow} S_n$
- A_{10}: $U_n| \equiv S_n \Rightarrow U_n \overset{SK}{\leftrightarrow} S_n$

Now with the help of BAN logic rules and assumptions, the proof of the proposed scheme will be performed.

From Msg 1, it will get that

S_1: $BS \lhd (UID, I)_{\text{h(UID} \| \text{K)}}$

Using assumption A_4 and message meaning rule, it will get that

S_2: $BS| \equiv U_n(UID, I)$

From Msg 2, it will get that

S_3: $BS \lhd (UID, SID, I, J)_{\text{ h(SID} \| \text{K)}}$

Using assumption A_6 and message meaning rule, it will get that

S_4: $BS| \equiv S_n(UID, I, SID, J)$

From Msg 3, it will get that

S_5: $U_n \lhd (UID, SID, I, J, U_n \overset{J}{\leftrightarrow} S_n)_{\text{ h(UID}\|\text{ k)}}$

Using assumption A_4 and message meaning rule, it will get that

S_6: $U_n| \equiv BS(UID, SID, I, J, U_n \overset{J}{\leftrightarrow} S_n)$

Using assumption A_3 and freshness conjugation rule, it will get that

S_7: $U_n| \equiv BS \equiv (UID, SID, I, J, U_n \overset{J}{\leftrightarrow} S_n)$

Now break the conjunctions to produce

S_8: $U_n| \equiv BS \equiv (U_n \overset{J}{\leftrightarrow} S_n)$

According to assumption A_7, apply the jurisdiction rule to obtain

S_9: $U_n| \equiv U_n \overset{J}{\leftrightarrow} S_n$

Now session key *SK* is a secret parameter could obtain

S_{10}: $U_n| \equiv U_n \overset{sk}{\leftrightarrow} S_n$ **(Goal 1 is achieved)**

According to Msg 5, it will get that

S_{17}: $U_n \lhd (UID, SID, I, J, U_n \overset{SK}{\leftrightarrow} S_n)_{SK}$

Using assumption S_{10} and message meaning rule

S_{18}: $U_n| \equiv S_n(UID, SID, I, J, U_n \overset{sk}{\leftrightarrow} S_n)$

According to assumption A_1, apply the freshness conjunction rule to obtain

S_{19}: $U_n| \equiv S_n \equiv (UID, SID, I, J, U_n \overset{sk}{\leftrightarrow} S_n)$

According to S_{19} and BAN logic rule to break conjunction to produce

S_{20}: $U_n| \equiv S_n \equiv (U_n \overset{sk}{\leftrightarrow} S_n)$ **(Goal 2 is achieved)**

From Msg 4, it will get that

S_{11}: $S_n \lhd (UID, SID, I, J, U_n \overset{I}{\leftrightarrow} S_n)_{h(SID\| k)}$

Using assumption A_6 and message meaning rule

S_{12}: $S_n| \equiv BS \sim (UID, SID, I, J, U_n \overset{I}{\leftrightarrow} S_n)$

According to assumption A_2, apply the freshness conjunction rule to obtain

S_{13}: $S_n| \equiv BS \sim (UID, SID, I, J, U_n \overset{I}{\leftrightarrow} S_n)$

Now applying BAN logic rule to break conjunction rule, it will get that

S_{14}: $U_n| \equiv BS \equiv (U_n \overset{I}{\leftrightarrow} S_n)$

Using assumption A_6 and jurisdiction rule, it will get that

S_{15}: $S_n| \equiv (U_n \overset{I}{\leftrightarrow} S_n)$

Now session key *SK* is a secret parameter could obtain

S_{16}: $S_n| \equiv (U_n \overset{sk}{\leftrightarrow} S_n)$ **(Goal 3 is achieved)**

According to Msg 6, it will get that

Table 6 Throughput measurement during encryption and decryption

Operations using Paillier system	Throughput (k bits/s)		
	n = 512	n = 1024	n = 2048
Encryption	98	28	8.2
Decryption	198	58	14.8

Table 7 Performance comparison of different algorithms

Authentication Scheme	Registration phase	Login phase	Authentication phase	Total
This Scheme	4 T_H	4T_H	4T_H	12 T_H
Dhillon and Kalra [30]	8 T_H	6 T_H	8 T_H	22 T_H
Xue et al. [31]	7 T_H	6 T_H	13 T_H	26 T_H
Turkanovi et al. [32]	7 T_H	5 T_H	7 T_H	19 T_H

Table 8 Functionality comparison of the scheme with existing protocols

Functionalities	Dhillon and Kalra [30]	Xue et al. [31]	Turkanovi et al. [32]	This scheme
Man-in-the-middle attack	Yes	No	Yes	Yes
Impersonation attack	Yes	No	No	Yes
Mutual authentication	Yes	Yes	Yes	Yes
Replay attack	Yes	Yes	Yes	Yes
Perfect forward secrecy	No	No	No	Yes
Insertion attack	No	No	No	Yes
Server spoofing attack	No	No	Yes	Yes
Password guessing attack	No	No	Yes	Yes

S_{21}: $S_n \lhd (UID, SID, I, J, U_n \overset{sk}{\leftrightarrow} S_n)$ SK

Using S_{19} and applying message meaning rule to produce

S_{22}: $S_n| \equiv U_n(UID, SID, I, J, U_n \overset{sk}{\leftrightarrow} S_n)$

According to assumption A_2 and freshness conjunction rule

S_{23}: $S_n| \equiv U_n \equiv (UID, SID, I, J, U_n \overset{sk}{\leftrightarrow} S_n)$

According to S_{23} to break conjunctions to produce

S_{24}: $S_n| \equiv U_n \equiv (U_n \overset{sk}{\leftrightarrow} S_n)$ (**Goal 4 is achieved)**

From Goal 1, Goal 2, Goal 3, and Goal 4 it shows that the session key shared by the user and BSN server is only known to them only.

5 Performance Analysis

Performance analysis of the proposed scheme has been done in following two parts:

A. The performance of the proposed protocol is measured in terms of computation cost (total time required for performing complex operation such as hashing) and communication cost (total no of bits transmitted). Also the throughput is measured for different modulus size shown below in Table 6.

The communication cost is mainly calculated for considering message transmission. Assuming that the identity (UID, SID) are 10 bytes long and the random nonce (N_1, N_2), secure one-way hash function are 160 bit long, the total no. of bits exchanged between user

Table 9 Remote data transfer delay for equal sample size

Data sample	Parameter range	Data arrival time (s)	Sample size (byte)	Data rate (kbs)	Delay (s)
ECG	2.5–3.0 mV.	0.003	2	0.50	1.031
Blood pressure	20–300 mm HG	0.02	2	1.2	1.135
Blood flow	1–300 ml/s	0.002	2	0.52	1.028

and base station is 940 bits (for transmitting msg 1, msg 3). Total no. of bits exchanged between base station and BSN server is 940 bits (for transmitting msg 2, msg 4). For mutual authentication total no. of bits exchanged is 976 bits (for transmitting msg 5, msg 6, msg 7, msg 8). Performance comparison is shown in Table 7 where T_H is the time required to perform one hashing.

Now the proposed scheme is compared with other existing algorithms in Table 8. Functionality Comparison shows that the scheme is favorably comparable with existing protocols.

B. To evaluate the performance of overall proposed system model, a database is considered from the hospital server which is mainly used for remote patient monitoring. The data samples are collected from different body sensors and aggregated in the hospital server into large packets to improve the transmission efficiency. The simulation scenario is considered that patient data (suppose ECG Signal, blood pressure and blood flow) is transmitted through the network coordinator to the remote computer (Doctor's laptop). OPNET simulator is used for this purpose. The total end to end delay is summarized in Table 9 given bellow:

The Table 9 shows that it is possible to transfer patient data within a reasonable delay for remote monitoring purpose or any emergency cases.

6 Conclusion

In this paper, a strong efficient authentication protocol for Wireless Body Sensor Networks is being proposed which can be applied in healthcare applications. It resist all possible attacks in distributed networks. The scheme provides mutual authentication between target server and user and also generate different session key for different server. The proposed scheme is based on Paillier cryptosystem which have homomorphic and self-binding property. For privacy preservation this type of system is very useful. The experiment results show that this scheme has low computational load and communicational load.

References

1. Samaneh, M., Mehran, A., Justin, L., David, S., & Abbas, J. (2014). Wireless body area networks: a survey. *IEEE Communications Surveys & Tutorials, 16*(3), 1658–1686.

2. Zimmerman, T. G. (1996). Personal area networks: Near-field intra body communication. *IBM Systems Journal, 35*(3/4), 609–617.
3. Shi, L., Li, M., Yu, S., & Yuan, J. (2012). "BANA: Body area network authentication exploiting channel characteristics. In *Proceedings of 5th ACM Conference Security. Privacy Wireless Mobile Network*, Tucson, AZ, USA: ACM, pp. 1–12.
4. Cai, L., Zeng, K., Chen, H., & Mohapatra, P. (2011). "Good neighbor: Ad hoc pairing of nearby wireless devices by multiple antennas. In *Proceedings of Network Distributed System Security Symposium*, pp. 1–15.
5. Shi, L., Yuan, J., Yu, S., & Li, M. (2013). "ASK-BAN: Authenticated secret key extraction utilizing channel characteristics for body area networks. In *Proceedings* of the *6th ACM Conference* on *Security* and *Privacy* in *Wireless* and *Mobile Networks*, pp. 155–166.
6. Shi, L., Li, M., Yu, S., & Yuan, J. (2013). BANA: Body area network authentication exploiting channel characteristics. *IEEE Journal on Selected Areas in Communications, 31*(9), 1803–1816.
7. Varshavsky, A., Scannell, A., LaMarca, A., & DeLara, E. (2007). "Amigo: Proximity-based authentication of mobile devices. In *Proceedings of 9th International Conference* on *Ubiq computing* (pp. 253–270). Berlin, Germany: Springer.
8. Kalamandeen, A., Scannell, A., DeLara, E., Sheth, A., & LaMarca, A. (2010). Ensemble: Cooperative proximity-based authentication. In *Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services* (pp. 331–344). New York, NY, USA: ACM.
9. Mathur, S., Miller, R., Varshavsky, A., Trappe, W., & Mandayam, N. (2011). Proximate: Proximity-based secure pairing using ambient wireless signals. In *Proceedings of the 9th International Conference on Mobile Systems, Applications, and Services* (pp. 211–224).
10. Poon, C., Zhang, Y., & Bao, S. (2006). A novel biometrics method to secure wireless body area sensor networks for telemedicine and mhealth. *IEEE Communications Magazine, 44*(4), 73–81.
11. Singh, K., Muthukkumarasamy, V. (2007). "Authenticated key establishment protocols for a home health care system. In *Proceedings of 3rd International Conference on Intelligent Sensors, Sensor Networks and Information (ISSNIP'07)* (pp. 353–358).
12. Venkatasubramanian, K., & Gupta, S. (2010). Physiological value based efficient usable security solutions for body sensor networks. *ACM Transactions on Sensor Network, 6,* 31:1–31:36.
13. ElGamal, T. (1985). A public key cryptosystem and a signature protocol based on discrete logarithms. *IEEE Transactions on Information Theory, 31*(4), 469–472.
14. Rivest, R., Shamir, A., & Adleman, L. (1978). A method for obtaining digital signatures and public key cryptosystems. *Communications of the ACM, 21*(2), 120–126.
15. Li, M., Yu, S., Lou, W., & Ren, K. (2010). Group device pairing based secure sensor association and key management for body area networks. In *Proceedings of IEEE INFOCOM* (pp. 1–9).
16. He, Debiao, Zeadally, Sherali, Kumar, Neeraj, & Lee, Jong-Hyouk. (2017). Anonymous authentication for wireless body area networks with provable security. *IEEE Systems Journal, 11*(4), 2590–2601.
17. Venkatasubramanian, K., Banerjee, A., & Gupta, S. (2010). Pska: Usable and secure key agreement scheme for body area networks. *IEEE Transactions on Information Technology in Biomedicine, 14*(1), 60–68.
18. Abi-char, P. E., Mhamed, A., & El Hassan, B "A secure authenticated key agreement protocol based on elliptic curve cryptography. In *International Symposium on Information Assurance and Security, IEEE* (vol. 57, pp. 89–94).
19. Bringer, J., Hervé, C., & Thomas I (2010) "Password based key exchange protocols on elliptic curves which conceal the public parameters. In *ACNS 2010, Lecture Notes in Computer Science* (vol. 6123/2010, pp. 291–308).
20. Chatterjee, K., De, A., & Gupta, D. (2015). A secure and efficient authentication protocol in wireless sensor network. *Wireless Personal Communications, 81*(1), 17–37.
21. Chatterjee, K., De, A., & Gupta, D. (2011). "Timestamp based authentication protocol for smart card using ECC. In *Proceedings of International Conference on Web Information System and Mining (WISM 2011),LNCS 2008* (vol. 6987, pp. 368–375).
22. Lim, M.-H., Yeoh, C.-M., Lee, S., Lim, H. & Lee, H. (2008). A secure and efficient three-pass authenticated key agreement protocol based on elliptic curves. In *Networking, LNCS 2008* (vol. 4982/2008, pp. 170–182).
23. Shamir, A. (1984). "Identity based cryptosystems and signature schemes. In *Proceedings of Advanced cryptology (CRYPTO'84)* (pp. 47–53). Berlin, Germany: Springer.
24. Yang, J., & Chang, C. (2009). An ID-based remote mutual authentication with keyagreement scheme for mobile devices on elliptic curve cryptosystem. *Computer Secuity, 28*(3–4), 138–143.

25. Yoon, E., & Yoo, K. (2009). "Robust ID-based remote mutual authentication with key agreement protocol for mobile devices on ECC. In *Proceedings of International Conference* on *Computer Science and Engineering*, Vancouver, Canada (pp. 633–640).
26. He, D., Chen, J., & Hu, J. (2012). An ID-based client authentication with key agreement protocol for mobile client-server environment on ECC with provable security. *Information Fusion, 13*(3), 223–230.
27. Wang, D., & Ma, C. (2013). Cryptanalysis of a remote user authentication scheme for mobile client-server environment with provable security based on ECC. *Information Fusion, 41*(4), 498–503.
28. Islam, S., & Biswas, G. (2011). A more efficient and secure ID-based remote mutual authentication with key agreement scheme for mobile devices on elliptic curve cryptosystem. *Journal of Systems and Software, 84*(11), 1892–1898.
29. Truong, T., Tran, M., & Duong, A. (2012). "Improvement of the more efficient and secure ID-based remote mutual authentication with key agreement scheme for mobile devices on ECC. In *Proceedings of the 26th International Conference Advanced Information Networking Application Workshops* (pp. 698–703).
30. Dhillon, P. K., & Kalra, S. (2017). A lightweight biometrics based remote user authentication scheme for IoT services. *Journal of Information Security and Applications, 34,* 255–270.
31. Xue, K., et al. (2013). A temporal-credential-based mutual authentication and key agreement scheme for wireless sensor networks. *Journal of Network and Computer Applications, 36*(1), 316–323.
32. Turkanović, Muhamed, Brumen, Boštjan, & Hölbl, Marko. (2014). A novel user authentication and key agreement scheme for heterogeneous ad hoc wireless sensor networks, based on the Internet of Things notion. *Ad Hoc Networks, 20,* 96–112.
33. Al-Janabi, S., et al. (2017). Survey of main challenges (security and privacy) in wireless body area networks for healthcare applications. *Egyptian Informatics Journal, 18*(2), 113–122.
34. O'Keeffe, M. (2008). The paillier cryptosystem. *Mathematics Department April, 18,* 1–16.
35. San, I., et al. (2016). Efficient paillier cryptoprocessor for privacy- preserving data mining. *Security and Communication Networks, 9*(11), 1535–1546.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Kakali Chatterjee is working as Assistant Professor in Computer Science and Engineering Department of National Institute of Technology Patna, India. She has completed her Ph.D. from Delhi University (formerly Delhi College of Engineering). She has involved in information security project of Govt. of India. She has many published research papers in LNCS (*Springer*) and reputed International Journals of eminent publishers like Springer and Elsevier. She is working in the field of Information Security and Cryptography for last 12 years.

Attention-guided deep convolutional neural networks for skin cancer classification

Arshiya Aggarwal
Delhi Technological University,
Delhi, India
Email: arshiya.aggarwal07@gmail.com

Nisheet Das
Delhi Technological University,
Delhi, India
Email: nisheetdas1@gmail.com

Indu Sreedevi
Delhi Technological University,
Delhi, India
Email: s.indu@dtu.ac.in

***Abstract*—Skin cancer is a silently killing disease which commonly goes unnoticed in its primitive stage but proves to be deadly later on. Hence, it needs to be detected and classified in the early stages itself. The advent of machine learning as well as deep learning based classification techniques has made this task possible. Deep convolutional neural networks (D-CNNs) have the ability to extract universal and dataset-specific features for the image classification task. But the classification of skin cancer images remains a challenging task due to the absence of balanced class images, difference between images of the same class, similarity between inter-class images and the inefficiency in focusing on the semantically significant areas of the image. To improve the performance of these D-CNNs, we incorporate the attention mechanism that focuses on the regions of importance in an image. In that regard, we propose an attention-guided D-CNN for classification of skin cancer. It is observed from the classification results that a model with attention boosts the accuracy of a normal D-CNN architecture by approximately 12%. Our research work contributes significantly to the field of biomedical image processing by providing a mechanism to improve performance of D-CNNs and facilitating early detection of skin cancer.**

***Index Terms*—skin cancer classification, deep learning, computer vision, dermoscopy, convolutional neural networks, attention, transfer learning, fine tuning**

I. Introduction

Due to the rapid depletion of the ozone layer and the increased exposure to the harmful UV radiations, skin cancer has become a predominant phenomenon, accounting for 40% of the total cancer cases worldwide [1]. Moreover, owing to its resemblance with common skin diseases and moles, it is generally neglected in its primitive phases. Due to its early stage negligence, the cancer cells do not remain isolated and spread out in the body. A quintessential approach for reducing the extent of mortality due to skin cancer is its early detection so that a thorough examination can be conducted and a timely treatment can be initiated. The advent of deep learning and the use of D-CNNs has helped achieve dermatologist level accuracy [2] to detect and point out the possibilities of skin cancer. Having given consideration to the criticality of this problem, we use skin cancer classification to illustrate the improvement in D-CNN architecture with the proposed attention mechanism.

978-1-7281-3975-3/19/$31.00 ©2019 IEEE

In this research work, we propose an attention-guided D-CNN for the classification of skin cancer into three major categories: Basal cell carcinoma (BCC), Melanoma and Squamous cell carcinoma (SCC). We have fine-tuned pre-trained D-CNN models that have the ability to extract universal features in images like edges, curves, textures etc, to help them learn dataset specific fine-grained features. Using attention helps the model to learn where to focus in the image by learning filters that activate the pixels of the important regions in the image. Our attention-guided model improves the accuracy of D-CNN by about 12%, is faster with lesser parameters and also eliminates the lengthy training procedures.

The rest of the paper is organized as follows: the next section talks about the methods implemented in the past literature in this field, followed by introduction to skin cancer. Further, the technology stack used for the research work has been introduced along with the proposed methodology. Finally, the paper is wrapped-up with results and conclusions that can be drawn from the conducted work.

II. Literature Review

Copious research work has been done in the field of skin cancer classification. Detection of skin cancer mainly consists of two tasks: feature extraction from the input image and thus, using them for classification purpose. Numerous techniques were devised for feature extraction purpose like ELM 7-point checklist, ABCD (asymmetry, border, color, diameter), Manzies method [3] and edge and contour detection (for initial analysis) [4]. Apart from that, filtering techniques like Laws masks, Gabor filter bank and wavelet transform [5] have been used for feature extraction. Use of deep learning methods reduces the efforts in feature extraction due to effectiveness of convolutional neural network (CNN) [6] [7] working in computer vision applications. Once the features are extracted, the next task is to choose the optimum features for classification. The classifiers used in the past are softmax classifier [8] , k-NN [9] [3] [10] (k-nearest neighbors), decision-making systems based on fuzzy logic systems (FLS) [4] and membership function theory. Also, artificial neural networks (ANNs) [9] [11] [3] [5], multi-class support vector machines (MCSVMs) [12], radial basis functions neural network (RBFNN) [3], adaptive wavelet based tree structure classifier (ADWAT) and classification and

regression decision trees (CART) [3] [13] have also been employed for classification purposes. Use of hybrid classifiers tends to improve feature selection and classification. [10] used hybrid k-NN and SVM to obtain better performance. The use of deep learning tools such as convolutional neural networks (CNNs) make the process easier as they combine the feature extraction and classification steps and can be trained end to end with efficient optimization techniques. Some recent research works have made use of the versatility of convolutional neural networks (CNN) with varied complex architectures to achieve better performances [2] [14] [15] [16]. In these works, the performance of the network has been boosted by using a pre-trained D-CNNs like GoogleNet/Inception v3, VGG16, VGG19 etc. Lately, a model using patch based attention has been introduced wherein high resolution images are divided into patch grids and are weighted according to their importance [17]. Our model makes use of pixel-based attention and out-performs the earlier proposed methods.

In our work, we incorporate the pixel wise attention mechanism to boost the performance of D-CNN. This helps the model to focus on the semantically important pixels in an image, improves the accuracy significantly, reduces the number of parameters and helps the model converge faster.

III. Skin Cancer

Cancer refers to the uncontrolled and abnormal growth of the specific type of cells in any part of the body. These cells could then spread to other body parts as well thus damaging them. Skin cancer is one such silently killing disease. There could be many reasons contributing to skin cancer, the most common one being exposure to UV radiations from the sun. Due to ozone layer depletion, the UV exposure has increased drastically over the last few years and so has the possibility of developing skin cancer. Other reasons could be genetic inclusion, light skin thickness or lack of immunity.

Majorly, skin cancer can be of two types namely melanoma skin cancer (MSC) and non-melanoma skin cancer (NMSC). Melanoma is a dangerous and very aggressive class of skin cancer which could be highly life threatening. Non-melanoma type majorly includes basal cell carcinoma (BCC) and squamous cell carcinoma (SCC). Non-melanoma types are generally not life threatening and can be cured by proper treatment methods such as a combination of surgery and chemotherapy. Early treatment of melanoma can also help in proper recovery and thus decreasing the mortality due to skin cancer. Table I describes the characteristic appearance and the main causes of the three major subdivisions of skin cancer. Figure 1 illustrates the sample images from our dataset.

TABLE I
Three major types of skin cancer

TYPE	CHARACTERISTICS	CAUSES
Basal-Cell Carcinoma (BCC)	Develops as small, round red spots which are slightly raised like a boil on skin.	• Excessive UV exposure • Lightly dense skins • High level exposure to arsenic • Poor immunity
Squamous-Cell Carcinoma (SCC)	Lumpy skin with hard shelly yellow coloured skin around.	• High UV radiation exposure • Poor immune system • Long staying wounds on skin • Light skin density
Melanoma	Appears as an extension of a mole with colour change, irregular shape and size and may also have itchiness.	• Low level of skin pigment making exposure easy • Heavy UV radiation exposure • Poor immunity

Fig. 1. Skin cancer category samples

IV. Proposed Methodology

We propose a novel attention guided D-CNN architecture for the classification of skin cancer. Our model is essentially a deep neural network with a pre-trained base model that is fine-tuned for the purpose of learning dataset-specific features followed by attention layers that aggregate the maximum activations from the feature maps generated and learn a filter to focus on the regions of importance of the image. The complete model has been segmented into the following parts to elucidate it in its entirety.

A. Convolutional Neural Networks (CNN)

Convolutional neural network (CNN) [18] [16] [7] is a deep learning algorithm used for feature extraction and classification. They have the ability to extract low-level as well as the high-level features omitting the minuscule noise from an image. Initially, the image is processed by performing convolutions using sliding window filters over the image. After the initial convolutional layers, low-level features such as edges, patterns, gradients etc. are obtained. Suitable padding is provided to the images to prevent loss of semantically important border information. The number of filters and the convolutional layers are decided based on the complexity of the problem. To aggregate semantically relevant information and to obtain the predominant features and activations, pooling is used. Pooling reduces the feature space dimensions and suppresses the noise. After feature extraction, decision-making is performed by the dense layers with softmax activation and categorical cross entropy loss. The dense layers are interspaced by dropout layers to get rid of over fitting. The classification is made using the predicted probabilities of the various classes. A CNN can thus, wholly perform feature extraction and classification, using extracted features.

B. Fine tuning D-CNN architectures

D-CNN's have millions of parameters. Training these architectures from scratch on datasets that have fewer images than the number of parameters affects the model's ability to generalize and leads to over fitting. Pre-trained models on massive diverse datasets like Imagenet capture universal features like curves, textures and edges in their early layers which are relevant to almost any image classification task, thus we freeze these weights and keep them intact. We need the model to learn dataset specific features by training the subsequent layers and changing the last layer. We used the following D-CNN architectures for our analysis:

1) InceptionV3: InceptionV3 [19] was the first runner-up in the 2015 ImageNet Large Scale Visual Recognition Competition (ILSVRC) with a top1 accuracy of 77.9% and top5 accuracy of 93.7%. Being developed by the Google Brain team, it has a size of 92 MB with 23,851,784 parameters. It's main characteristic is the Inception module which concatenates feature maps generated by kernels of varying sizes. The schematic diagram for the model has been shown in Figure 2.

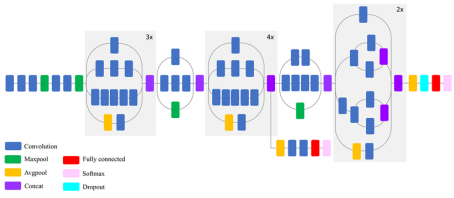


Fig. 2. InceptionV3 architecture

2) Xception: Xception [20] means the extreme version of Inception and was also developed by Google. It performs better than InceptionV3 on both ImageNet ILSVRC and JFT datasets. With the added depth wise separable convolution, this model has 22,910,480 parameters and a size of 88 MB. It achieves a top1 accuracy of 79% and a top5 accuracy of 94.5% on the ImageNet dataset. The schematic diagram for the model has been illustrated in Figure 3 .

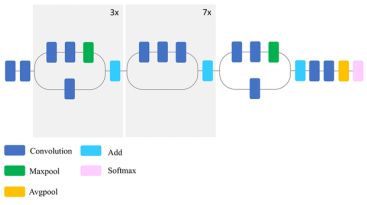


Fig. 3. Xception architecture

3) VGG16: VGG16 D-CNN architecture [21] (named after the Visual Geometry Group from Oxford) was the winner of ILSVRC 2014. It achieves a top1 accuracy of 71.3% and a top5 accuracy of 90.1% on the ImageNet dataset. It contains 138,357,544 parameters and is 528 MB large. It is a very heavy model in terms of the number of parameters but still is used by many deep learning practitioners owing to its simplicity. The schematic diagram of the model has been shown in the Figure 4.

Fig. 4. VGG16 architecture

C. Attention in images

Attention mechanism in CNN's (also called neural attention) helps to pay attention on a subset of the inputs. For the case of images, this mechanism helps the model to choose certain regions of an image that are more significant to the classification results than the rest of the regions. The fundamental procedure for any neural network classification task is a chain of matrix multiplications followed by an element wise activation function filter. Now, the attention mechanism tries to compute a mask that aims to act as an activation function to "switch on" the significant region pixels and "switch off" the others. Neural networks are universal function approximators. In attention, we aim to estimate a filter that can help enhance the regions of interest leaving out the irrelevant regions of the image [22] [23]. Let the input dimension be "d" and the feature space dimension at any point be "k". Let $x \in \mathbb{R}^d$ be the input vector, $z \in \mathbb{R}^k$ be a feature vector and $a \in [0,1]^k$ be the attention vector. Then, Equation (1) gives the attention vector "a" and the attention visualization "g", where ϕ denotes the parameters of the attention network.

$$\begin{aligned} a &= f_\phi(x) \\ g &= a \odot z \end{aligned} \tag{1}$$

The feature vector "z" has been computed by some other neural network such that $z = f_\theta(x)$ where θ denotes the parameters of this neural network. The vector "g" gives the result computed after applying the attention mask to the feature vector. This can be used for further analysis or prediction.

In our proposed architecture, the attention network has been incorporated as an additional branch in the main neural network. The feature maps that have been extracted from the preceding layers of the architecture contains a lot of information about the different characteristics of the image from the various regions of the image. The activations from the multiple feature maps are aggregated to form one map or filter using a separate convolutional neural network that acts as a function approximator for the filter. The feature maps from this separate branch learn the relative importance (or weight) of the different region pixels through their various layers and activate the regions of importance as opposed to the others. These feature maps are clipped between [0,1]. This clipped feature map acts as the mask to switch on or off a particular pixel in a particular feature map. This

is fanned out across all the channels and an element wise multiplication is performed between the masks and the initial feature maps to properly implement the attention mechanism. This element wise multiplication gives the different regions in all the initially extracted feature maps a weight in [0,1] based on their relative importance in the classification task. This process has been illustrated in the Figure 5 where N denotes the number of examples, Z the number of feature maps, X the height and Y the width of the image. After deliberate experimentation, the following hyper-parameters (including the number of feature maps) were chosen so as to extract the best performance results from the model.

Fig. 5. Attention mechanism for the proposed model

The inclusion of this mechanism helps the model perform much better. This branch helps in estimating the importance of the different regions in the image. It also helps the model to understand the various regions in the image better by extracting information about what the region signify and how that region contributes to the classification task. Based on this information, the model trains the weights of this branch to compute a filter to focus on the important regions in an image. This branch improves the performance of D-CNN thus, proving to be a significant contributor in this field.

D. Implementation details

In our work, we propose an attention guided deep convolutional neural network architecture for the classification of skin cancer into the three major categories: Basal Cell Carcinoma (class 0), Melanoma (class 1), Squamous Cell Carcinoma (class 2). The training and testing datasets each contain more than 1000 images per class that have been obtained from multiple public datasets including HAM10000 [24], Dermanet, BioGPS, ISIC images and Google Images. Images downloaded from Google Images and other sources were stored separately for the three classes, loaded and labeled. To reduce any imbalance in the dataset data augmentation techniques like translation, rotation and horizontal flip have been used in limited amount on the images. To further reduce the imbalance, class weights for the various classes in proportion to the class size have been used while compiling the model. The dataset images are then shuffled and normalized in [0,1] to avoid any unwanted bias. The pre-trained base models are loaded, the starting layers are frozen and the subsequent last few layers are made trainable. Further on, the attention mechanism is implemented to help aggregate information about the relative semantic importance of the various regions from the initially generated feature maps. The attention-weighted feature maps are generated by performing an element wise multiplication between the feature maps (generated using D-CNN) and the attention filter. These attention-weighted feature maps are fed to dense layers to make a prediction using the softmax activation function. Batch normalization and dropout layers are used to prevent over fitting. The proposed architecture has been diagrammatically illustrated in Figure 6.

V. Simulation Results

For the training process, k-fold cross validation is used with k=5 so as to utilize the data well. Also, the batch size is increased and learning rate is decreased with increasing epochs to avoid over fitting and help the model converge better. After the model is trained, the attention maps are plotted as shown

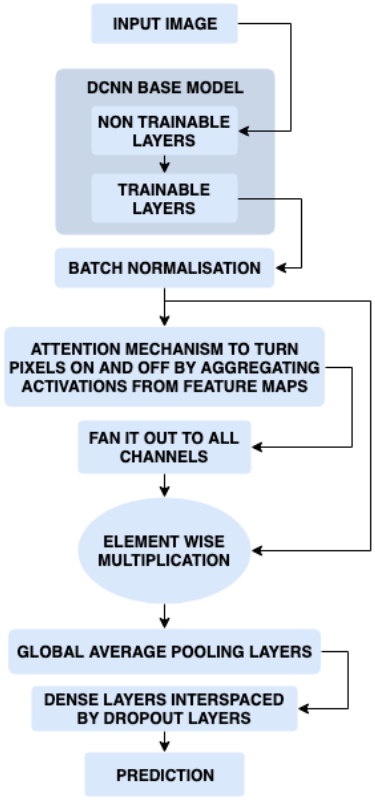


Fig. 6. Flowchart for the proposed model

in Figure 7. It can be observed from the attention maps that the model is able to understand the important regions in the image.

The proposed model is compiled using the Adam optimizer and categorical cross entropy loss, using Equation (2).

$$\text{loss}(p^{\text{true}}, p^{\text{predict}}) = -\sum_i p_i^{\text{true}} \log p_i^{\text{predict}} \tag{2}$$

The performance of the model is analyzed using various metrics like accuracy (Equation (3)), precision (Equation (4)), recall (Equation (5)) and F1 score (Equation (6)).

$$\text{accuracy}\left(p^{\text{true}}, p^{\text{predict}}\right) = \frac{\text{argmax}(p^{\text{true}}) == \text{argmax}(p^{\text{predict}})}{\text{total examples}} \tag{3}$$

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}} \tag{4}$$

$$\text{Recall} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}} \tag{5}$$

$$\text{F1 score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

While accuracy gives an idea of the percentage of examples predicted correctly, we cannot solely rely on it as it does not give sufficient information. Hence, we use other metrics as well. Precision gives an idea about the number of examples that were predicted as a particular class and belonged to that class as opposed to the total examples predicted as that class. Recall, on the other hand gives an idea about the number of examples that were predicted as a particular class and belonged to that class as opposed to the total examples that belonged to that class. These metrics are a better way to evaluate a classification model as compared to accuracy. The F1 score combines these two metrics and is used when a balance of precision and recall is required. The classification results for the three D-CNN architectures with and without attention have been recorded in Table II. It is evident from the results that there is great improvement in the scores when attention is incorporated. Amongst the three D-CNN architectures, InceptionV3 has the best overall results. VGG16 has somewhat comparable results but due to the large number of parameters and increased computational complexity, InceptionV3 is preferred as the base pre-trained model.

To further evaluate the performance of the model we resort to the visual metric, confusion matrix. Confusion matrix shows the actual classes on the vertical axis and the predicted classes on the horizontal axis. It helps visualize the performance of the model vividly and judge the model critically.

Fig. 7. Attention maps

Fig. 8. Confusion Matrices

From the confusion matrices in the Figure 8 it is clear that InceptionV3 base model performs better than the other two models as the dark blue shade belongs to only the diagonal elements meaning maximum number of samples are predicted within their own class.

Thus, after performing an exhaustive classification evaluation we conclude that incorporating attention improves the classification results significantly. This is because the convolutional layers of the attention branch help the model aggregate the important information from the feature maps and better understand the significance of the various regions in the image. This aggregation is used to learn an attention filter which is in turn used to generate the attention-weighted feature maps by performing an element wise product with the feature maps generated from the D-CNN. Moreover, amongst the models with attention InceptionV3 performs the best on our dataset.

VI. Conclusion

In the presented work, the issue of early stage negligence of skin cancer has been addressed by devising an algorithm to improve the classification performance of D-CNNs using an attention guided deep convolutional neural network. The various types of skin cancer along with their characteristics and causes have been described along with a brief description of the previous works. The use of pre-trained D-CNN architectures helps the model converge to minimum loss within a few epochs thus, eliminating the long training procedures. The attention mechanism has been used in a D-CNN model which improves its performance significantly. The performance of

TABLE II
CLASSIFICATION RESULTS

S NO.	MODEL	ACCURACY	PRECISION	RECALL	F1 SCORE
1	**INCEPTIONV3 +ATTENTION (PROPOSED MODEL)**	**85.8%**	**0.87**	**0.86**	**0.86**
2	XCEPTION +ATTENTION	83.6%	0.86	0.83	0.83
3	VGG16 +ATTENTION	85.2%	0.87	0.85	0.86
4	INCEPTIONV3	72.8%	0.74	0.72	0.73
5	XCEPTION	71.3%	0.76	0.65	0.7
6	VGG16	74.7%	0.79	0.67	0.73

three D-CNN architectures with and without attention have been recorded and tabulated. There is an improvement of around 12% in the classification accuracy for the models with attention. This is because the attention mechanism computes a mask which assigns a relative weight to the various regions in the extracted feature maps thus helping them to focus on the important parts of the image. It is also observed that InceptionV3 performs better in terms of classification results and computational complexity than the other D-CNN's. Thus, the model **InceptionV3 + attention** proves to be efficient in terms of classification scores, training time and number of parameters.

For the future, better hyper parameters can be employed for the attention branch to boost its performance further. Larger models with more attention blocks can be used to produce better masks for the skin cancer images which might improve the classification accuracy. Better transfer learning approaches can be used [25] wherein all convolutional layer weights are fine-tuned. Balanced curated datasets in this area can be produced and made available publicly as the scope in this field is growing at an exponential rate. Moreover, a faster and efficient model can be proposed to separate the different classes distinctly. The base models can first be trained on a dataset related to skin diseases so as to ease out fine-tuning on skin cancer datasets.

The research work contributes in improving the effectiveness of D-CNNs in early stage cancer detection by incorporating the attention mechanism. This research work makes use of the advance technologies to bring about a change in the current scenario of deaths caused by skin cancer and thus help the world's population be cautious and healthy.

REFERENCES

[1] *National Library of Australia Cataloguing–in–Publication data: Lifestyle and cancer: knowledge, attitudes and behavior in NSW 2009 SHPN (CI) 120203*, 2012.

[2] A. Esteva, B. Kuprel, R. A. Novoa, K. J., S. M. Swetter, H. M. Blau, and S. Thrun, ""Dermatologist-level classification of skin cancer with deep neural networks", Nature," pp. 542–115, 2017.

[3] S. R. . H. AN, *,"Review on automatic early skin cancer detection.'*, Nanjing, China, 6 2011, vol. 2011.

[4] J. Blackledge and D. Dubovitskiy, "Object detection and classification with applications to skin cancer screening." *ISAST Transactions on Intelligent Systems*, vol. 1, no. 1, pp. 34–45, 2008.

[5] "SVM-based texture classification and application to early melanoma detection." *, 28th annual international conference of the IEEE Engineering in Medicine and Biology Society*, pp. 4775–4778, 0.

[6] . L. H, "A deep learning approach to universal skin disease classification." 2016.

[7] . E. Nasr-Esfahani, *Melanoma detection by analysis of clinical images using convolutional neural network,*, Orlando, FL, 2016.

[8] "A deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection," 2013, pp. 8150–403.

[9] . M. Elgamal, "Automatic Skin Cancer Images Classification," *,International Journal of Advanced Computer Science and Applications(IJACSA)*, vol. 4, no. 3, p. 2013. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2013.040342

[10] D. G. . R. Sumithra, *Segmentation and Classification of Skin Lesions for Disease Diagnosis*, 2015, vol. 45. [Online]. Available: https://doi.org/10.1016/j.procs.2015.03.090.

[11] A. M. B, B. R, and H. F, ","Skin cancer recognition by using a neuro-fuzzy system."," *Cancer Inform*, vol. 10, no. 1-11, 2 2011.

[12] R. Sumithra and D. S. G. M. Suhil, "Segmentation and Classification of Skin Lesions for Disease Diagnosis, Procedia," *Computer Science*, vol. 45, pp. 76–85, 2015.

[13] "Maglogiannis I, Doukas CN, "Overview of advanced computer vision systems for skin lesions characterization."," *IEEE Trans Inf Technol Biomed*, vol. 5, no. 721–733, p. 9.

[14] U. O. Dorj, K. K. Lee, and J. Y. Choi, "The skin cancer classification using deep convolutional neural network," *Multimed Tools Appl*, vol. 77, pp. 9909–10, 2018.

[15] K. J. . Premaladha, ""Novel Approaches for Diagnosing Melanoma Skin Lesions Through Supervised and Deep Learning Algorithms", Med Syst," pp. 96–10, 2016.

[16] A. B. . J. Kawahara and G. Hamarneh, *Deep features to classify skin lesions,*, 2016.

[17] N. Gessert, T. Sentker, F. Madesta, R. Schmitz, H. Kniep, I. Baltruschat, R. Werner, and A. Schlaefer, "Skin Lesion Classification Using CNNs with Patch-Based Attention and Diagnosis-Guided Loss Weighting," *IEEE Transactions on Biomedical Engineering*, 2019.

[18] H. Liao, "A deep learning approach to universal skin disease classification," 2016.

[19] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[20] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[22] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.

[23] W. Köpp, P. van der Smagt, and S. Urban, "A Differentiable Transition Between Additive and Multiplicative Neurons," *arXiv preprint arXiv:1604.03736*, 2016.

[24] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific data*, vol. 5, p. 180161, 2018.

[25] S. Chaabouni, J. Benois-Pineau, and C. B. Amar, "ChaboNet: Design of a deep CNN for prediction of visual saliency in natural video," *Journal of Visual Communication and Image Representation*, vol. 60, pp. 79–93, 2019.

Carbon Nanotube Recessed Channel (CNT-RC) MOSFET for High Linearity/ULSI Applications

Ajay Kumar
Electronics and Communication Engineering Department
Jaypee Institute of Information Technology, Noida, India
ajaydce@ieee.org

Neha Gupta
Applied Science and Humanities Department
Dr. Akhilesh Das Gupta Institute of Technology and Management (IPU)
Delhi, India
gupta.neha@niecdelhi.ac.in

Samarth Singh
Applied Physics Department
Delhi Technological University
Delhi, India
samarthsingh_2k14@dtu.ac.in

Balark Tiwari
Applied Physics Department
Delhi Technological University
Delhi, India
balarktiwari_2k14@dtu.ac.in

Madan Mohan Tripathi
Electrical Engineering Department
Delhi Technological University
Delhi, India
mmtripathi@dce.ac.in

Rishu Chaujar
Applied Physics Department
Delhi Technological University
Delhi, India
chaujar.rishu@dtu.ac.in

***Abstract*— This work proposes, Carbon Nanotube (CNT) based Recessed Channel (RC) MOSFET for high linearity applications. CNT-RC-MOSFET reflected high linearity than Conventional (C) MOSFET and Conventional Recessed Channel (CRC) MOSFET in terms of higher drain current and transconductance (g_m) and reduced higher order of transconductance (g_{m2} and g_{m3}). Moreover, higher value VIP2, VIP3, IIP3 and lesser values of the IMD3 reflects high linearity in CNT-RC device as compared to its counterparts at sub-20 nm regime.**

Keywords— *CNT-RC-MOSFET, VIP2, VIP3, IIP3, IMD3, Linearity.*

I. INTRODUCTION

In this era of RFIC and wireless communication, one needs to achieve faster data exchange with lesser or no error possible [1]. Sources to these errors lie in signal noise, distortion and bad processing of the signal by a device which one can overcome these by the help of CMOS technology which provides low intermodulation distortion, thus maintaining linear device operations [2]. CNT-RC MOSFET is more promising over CRC and Conventional MOSFETs as CNT has exceptionally good electrical properties due to the unique electronic structure of graphene [3]. By controlling the diameter and chirality of CNT, we can adjust the bandgap, thus proving CNT-RC MOSFET to be a promising candidate in the nano-regime. Along with bandgap control, CNT benefits in better control over channel formation, threshold voltage, higher electron mobility, higher current density, and most saliently higher linearity.

In RF systems, linearity assessment is of much importance as it helps in keeping the errors caused by inter-modulation and high order harmonics to be at least as possible. Although it can be taken care by system-level techniques, but it demands complex circuitry. Other major obstacle is while downsizing the device (MOSFETs) some undesirable effects arises, curbing operational parameters such as DIBL (Drain Induced Barrier Lowering), SCE (Short Channel Effects) [4] etc. To tackle those problems, advancements in gate engineering have landed some pretty good options which not only provide immunity to these but also enhance the functionality of the device. Various gate engineered devices come into consideration from which

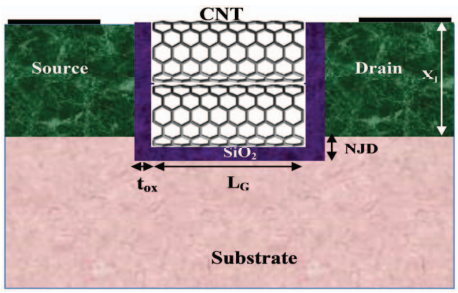


Fig. 1. Structure of CNT-RC-MOSFET. Here, t_{ox} = 2 nm, and L_G = 20 nm and L_{ch} (channel length)=44 nm. Source/drain regions are uniformly doped (5×10^{19} cm^{-3}, an n-type impurity) and substrate region is doped by p-type impurity (5×10^{16} cm^{-3}). NJD=10 nm while groove depth is 38 nm. V_{gs} = 1.5 V and V_{ds} = 0.2 V.

-CNT-RC MOSFET (Recessed Channel MOSFET) is considered in this paper for complete linearity analysis [1, 5] in terms of high power of transconductance and figure-of-merit (FOM) metrics, which are g_{m1}, g_{m2}, g_{m3} as first, second and third orders of transconductance and VIP2, VIP3, IIP3 as a second and third order of intercept points and IMD3 as third-order of intermodulation distortion. Along with comparison of CRC-MOSFET and conventional MOSFET being compared for better linearity analysis and device operations.

II. DEVICE PRINCIPLE, STRUCTURE AND SIMULATION METHODOLOGY

The device proposed in this paper is CNT-RC MOSFET (Carbon Nanotube Recessed Channel MOSFET) as described in Fig. 1, which prove to tackle SCEs and DIBL effects also maintaining the series resistance thus optimal for CMOS ULSI circuits. High-density electric field lines concave device structure resulting in the formation of two potential barriers at corners, consequently reducing punch through improving SCEs and DIBL [6]. There is a boundary around our gate region with an oxide layer, preventing direct contacts in between source & gate, Gate & drain and even gate 1 & gate two which in turns helps us to bear the effects caused by tunneling and leakage current. All simulations are done on Silvaco TCAD [7] using optimum models which include Lombardi CVT, Fermi Dirac, Shockley-Read-Hall (SRH) Recombination model, etc. to consider real-world possibilities.

978-1-7281-1895-6/19/$31.00 ©2019 IEEE

III. Device Calibration

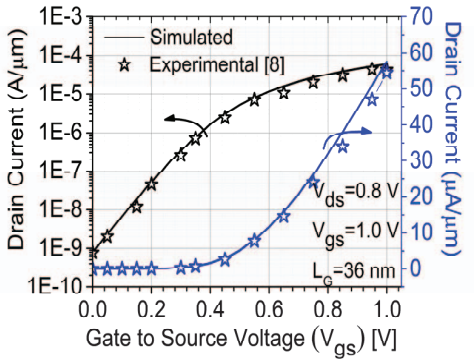
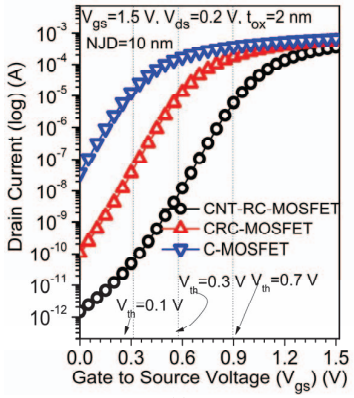


Fig. 2. Calibration of recessed channel MOSFET with experimental data of 36 nm.

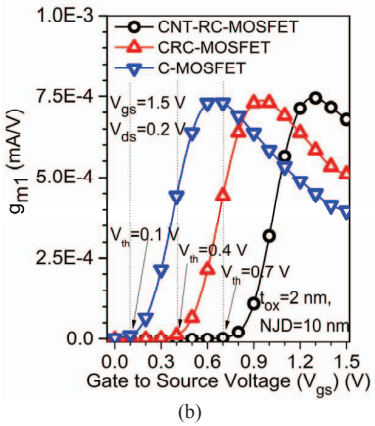
Further, the simulated device has been validated with the calibration of models with experimental data [8]. A fabricated 36 nm gate length grooved MOSFET with its experimental data [8] and a simulated device with same dimentions have been taken for the calibration plotted (as shown in Fig. 2).

IV. Results and Discussion

Transfer characteristics, as shown in Fig. 2(a) describes the nature and quality of a device very well, how better is the performance, turn ON response [9], etc.

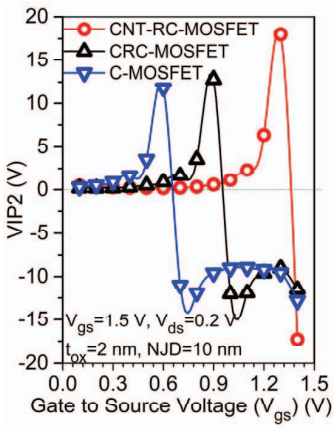


(a)

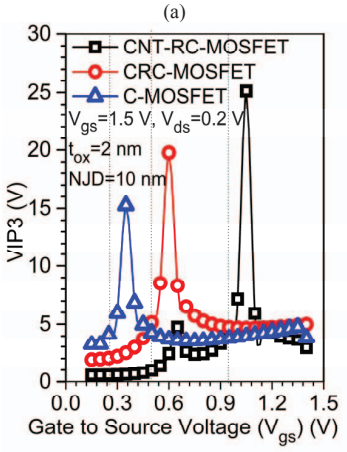


(b)

Fig. 3: (a) Drain current and (b) g_m variations with V_{gs} for CNT-RC, CRC, and C MOSFET-.



(a)



(b)

Fig. 4: (a) VIP2 and (b) VIP3 variations with V_{gs} for CNT-RC, CRC, and C-MOSFET.

When a device is scaled down, several effects originate which results in high distortion and making a device non-linear. If a device is non-linear and distorted, it will deplete more power.

Higher values of a higher order of transconductance [4, 10] coefficients, i.e. g_{m1}, g_{m2}, g_{m3}, assure higher linearity of a device. Transconductance is defined as the ratio of change in current at output electrode to the change in input voltage and is calculated as:

$$g_{mn} = \frac{1}{n!}\frac{\delta^n I_{ds}}{\delta V_{gs}^n}, \text{ where n = 1, 2, 3} \qquad (1)$$

Using the above relation, Fig. 3(b), 7(a) and 7(b) are plotted for g_{m1}, g_{m2}, and g_{m3}. From these plots, CNT-RC MOSFET holds higher values as compared to CRC and conventional MOSFETs as we go higher in order of transconductances. For optimum device operations, one should operate at Zero Cross over the point of g_{m3}. From Fig. 7(b), CNT-RC MOSFET has the lowest value. Thus it can be operated at lower gate voltage for best device operation. FOM matrices are the solution for distortion analysis, values of VIP2, VIP3, IIP3, and IMD3 can provide a clear picture about distortion state of the device [11, 12]. VIP2 is the extrapolated gate voltage at which first-order harmonics matches with second-order harmonics.

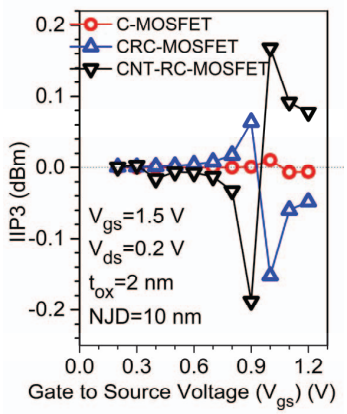


Fig. 5: IIP3 variations with V_{gs} for CNT-RC, CRC, and C-MOSFET.

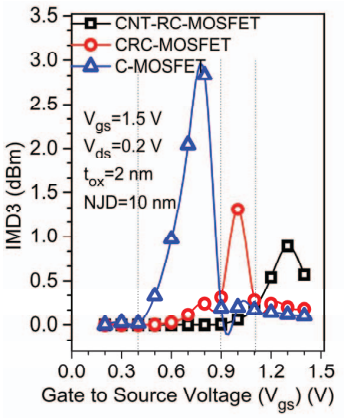


Fig. 6: IMD3 variations with V_{gs} for CNT-RC, CRC, and C-MOSFET.

Similarly, V_{IP3} is the extrapolated gate voltage where first order matches with third order harmonics. I_{IP3} is the extrapolated input power where first-order harmonics equalizes third order harmonics. The device FOMs can be evaluated as [5]:

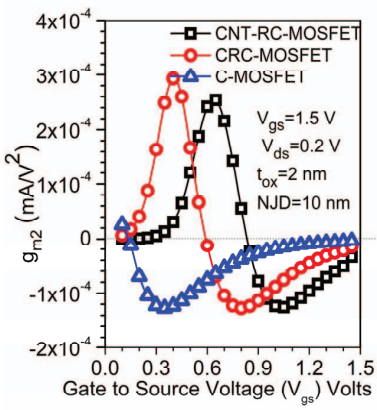
$$VIP2 = 4\frac{g_{m1}}{g_{m2}} \tag{2}$$

$$VIP3 = \sqrt{24 \times g_{m1} / g_{m3}} \tag{3}$$

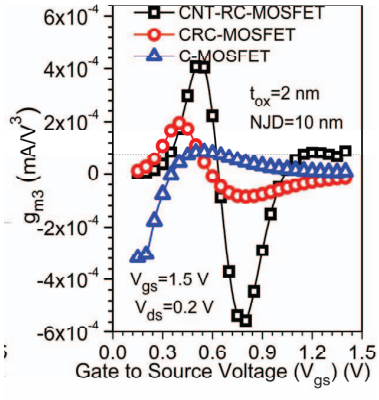
$$IIP3 = \frac{2}{3}\frac{g_{m1}}{g_{m3} \times Rs} \tag{4}$$

$$IMD3 = \left(\frac{9}{2} \times (VIP3)^3 \times g_{m3}\right)^2 \times R_S \tag{5}$$

Most of the cases for RFIC and wireless communications, R_s is taken as 50 Ω. For a device to be less or minimally distorted it should have high values of VIP2, VIP3, and IIP3 with low values of IMD3. From Fig. 3 (a), 3 (b), 3 (c) and 3 (d), CNT-RC MOSFET is a less distorted device as it has higher peaks of VIP2, VIP3, and IIP3 with low values of IMD3 in graphs plotted.



(a)



(b)

Fig. 7: (a) IMD3, (b) g_{m2}, and (c) g_{m3} variations with V_{gs} for CNT-RC, CRC, and C-MOSFET.

V. Conclusion

The present work concentrates on linearity and distortion analysis of CNT-RC MOSFET thus making it more promising than CRC and conventional MOSFETs since it has higher values of a higher order of transconductance, FOMs (VIP2, VIP3, and IIP3) and low value of IMD3. With its sub-20 nm design, it helps in developing ULSI applications that includes RF devices, etc.

References

[1] A. Kumar, "Effect of trench depth and gate length shrinking assessment on the analog and linearity performance of TGRC-MOSFET," *Superlattices and Microstructures,* vol. 109, pp. 626-640, 2017/09/01/ 2017.

[2] T. Y. Yum, L. Chiu, C. H. Chan, and Q. Xue, "High-efficiency linear RF Amplifier-a unified circuit approach to achieving compactness and low distortion," *IEEE transactions on Microwave Theory and Techniques,* vol. 54, pp. 3255-3266, 2006.

[3] S. Oh, S. W. Lee, D. Kim, J.-H. Choi, H.-C. Chae, S. M. Choi, *et al.*, "Low Power Switching Characteristics of CNT Field Effect Transistor Device with Al-Doped ZrHfO2 Gate Dielectric," *Journal of Nanomaterials,* vol. 2018, 2018.

[4] A. Kumar, N. Gupta, and R. Chaujar, "TCAD RF performance investigation of Transparent Gate Recessed Channel MOSFET," *Microelectronics Journal,* vol. 49, pp. 36-42, 2016.

[5] A. Kumar, M. Tripathi, and R. Chaujar, "Reliability Issues of In2O5Sn Gate Electrode Recessed Channel MOSFET: Impact of Interface Trap Charges and Temperature," *IEEE Transactions on Electron Devices,* vol. 65, pp. 860-866, 2018.

[6] A. Kumar, N. Gupta, and R. Chaujar, "Analysis of novel transparent gate recessed channel (TGRC) MOSFET for improved analog behaviour," *Microsystem Technologies,* vol. 22, pp. 2665-2671, 2016.
[7] I. SILVACO, "ATLAS User's Manual," *Santa Clara, CA, Ver,* vol. 5, 2011.
[8] J. Appenzeller, R. Martel, P. Avouris, J. Knoch, J. Scholvin, J. A. del Alamo*, et al.*, "Sub-40 nm SOI V-groove n-MOSFETs," *IEEE Electron Device Letters,* vol. 23, pp. 100-102, 2002.
[9] S. Sahay and M. J. Kumar, "Spacer design guidelines for nanowire FETs from gate-induced drain leakage perspective," *IEEE Transactions on Electron Devices,* vol. 64, pp. 3007-3015, 2017.
[10] A. Kumar, N. Gupta, and R. Chaujar, "Power gain assessment of ITO based Transparent Gate Recessed Channel (TGRC) MOSFET for RF/wireless applications," *Superlattices and Microstructures,* vol. 91, pp. 290-301, 3// 2016.
[11] N. Gupta and R. Chaujar, "Optimization of high-k and gate metal workfunction for improved analog and intermodulation performance of Gate Stack (GS)-GEWE-SiNW MOSFET," *Superlattices and Microstructures,* vol. 97, pp. 630-641, 2016.
[12] N. Gupta and R. Chaujar, "Investigation of temperature variations on analog/RF and linearity performance of stacked gate GEWE-SiNW MOSFET for improved device reliability," *Microelectronics Reliability,* vol. 64, pp. 235-241, 2016.

Materials Research Express

ACCEPTED MANUSCRIPT

Characterization and parametric optimization of % change in residual stress of Mo composite coating using Taguchi approach

To cite this article before publication: Ankit Tyagi *et al* 2019 *Mater. Res. Express* in press https://doi.org/10.1088/2053-1591/ab6252

Manuscript version: Accepted Manuscript

Accepted Manuscript is "the version of the article accepted for publication including all changes made as a result of the peer review process, and which may also include the addition to the article by IOP Publishing of a header, an article ID, a cover sheet and/or an 'Accepted Manuscript' watermark, but excluding any other editing, typesetting or other changes made by IOP Publishing and/or its licensors"

This Accepted Manuscript is **© 2019 IOP Publishing Ltd**.

During the embargo period (the 12 month period from the publication of the Version of Record of this article), the Accepted Manuscript is fully protected by copyright and cannot be reused or reposted elsewhere.
As the Version of Record of this article is going to be / has been published on a subscription basis, this Accepted Manuscript is available for reuse under a CC BY-NC-ND 3.0 licence after the 12 month embargo period.

After the embargo period, everyone is permitted to use copy and redistribute this article for non-commercial purposes only, provided that they adhere to all the terms of the licence https://creativecommons.org/licences/by-nc-nd/3.0

Although reasonable endeavours have been taken to obtain all necessary permissions from third parties to include their copyrighted content within this article, their full citation and copyright line may not be present in this Accepted Manuscript version. Before using any content from this article, please refer to the Version of Record on IOPscience once published for full citation and copyright details, as permissions will likely be required. All third party content is fully copyright protected, unless specifically stated otherwise in the figure caption in the Version of Record.

View the article online for updates and enhancements.

This content was downloaded from IP address 14.139.251.100 on 22/12/2019 at 07:18

Characterization and parametric optimization of % change in residual stress of Mo composite coating using Taguchi approach

Ankit Tyagi [a], SM Pandey[a], R S Walia[a, b], Qasim Murtaza[a]

[a] Department of Mechanical Engineering, Delhi Technological University, Delhi, India.

[b] Department of Production and Industrial Engineering, Punjab Engineering College, Chandigarh, India.

Abstract

The aim of present work was principally focus upon development of Mo composite coating using HVOF thermal spray coating process and to enhance residual stresses using Taguchi approach. The process parameter used in Taguchi experiment are Temperature, load and frequency. ANOVA method was used to examine and analyze the process parameters significance of Mo composite coating. The HRXRD and FESEM attached EDS were used to confirm surface morphology and presence of Mo based composite coating on steel substrate. Experimental results also concluded that for percent (%) change in residual stress, temperature (64.19 %) has maximum % contribution followed by frequency (21.82 %) and load (6.31 %).

Keywords: HVOF thermal spray coating, percent (%) change in residual stress, Taguchi optimization, Analysis of variance (ANOVA).

*Corresponding author.

E-mail address: waliaravinder@yahoo.com, qasimmurtaza@dce.ac.in, shaileshmanipandey@dce.ac.in

1. Introduction

Nowadays, strict rule and regulations aimed at environmental protection [1, 2] along with technical and economical need put limits in commonly used tribosystem in engines. The tribological behavior of IC engines (performance of piston rings) have significant effect concerning fuel consumption. A research study reveals that about 35-45% frictional losses IC engines are because of piston cylinder assembly [3]. Major losses in an automobile are because of friction and wear, which accounts for almost 33% of the total fuel energy. Significant efforts have been put in developing surface texturing, anti-wear and low friction coatings having good adhesion strength, high thermal conductivity and excellent tribological properties. Generally, in order to increase service life, $MoSi_2$, MoO, thermal spray Mo_2C, NiCrBSi, Chrome coated [4], CrC, NiCr and Al-Si alloy with Mo powder [5-10], Fe based coating, WC-CoCr, and Cr3C2-WC-NiCoCrMo [7], coatings were developed on the surface

Accepted Manuscript

of piston-cylinder. However, these coatings usually possess low thermal stability, and high internal stresses on the counter surfaces. Therefore, hard coating uses are diminishing in the piston cylinder, due to its limitations. Nowadays, carbon based coating like thermal CVD diamond, GLC, DLC [11] and doped DLC films of silicon, hydrogen, tungsten and so on, have been widely used for better tribological properties and are commercially accessible for the application in small volume commercial engines. Therefore, developing a new tribological coating for piston cylinder arrangement is a need for modern automobile industry. Taguchi approach is used to enhance the quality of goods and detect the control factors. It is also used to establish a relation between input variable and output variables. In Taguchi approach, DOE is important in order to design and investigate the desired response with statistical as well as mathematical model [12-15]. ANVOVA method was used to estimate the results obtained after DOE. It is also used to identify the effect of input parameters in the response. A study based on Taguchi coupled with ANN was used to study the optimal factors and result shows that ANN predict 98.4% R value on compared with regression model [14].

The main objective of present work was principally focuses upon developing Mo composite coating using HVOF thermal spray coating process and to enhance percentage (%) change in residual stress using Taguchi approach. The process parameter used in Taguchi experiment are Temperature, load and frequency. ANOVA method was used to examine and analyze the process parameters significance of Mo composite coating. The main objective is to develop carbon powder, CrC, Mo, and NiCr based composite coating having good surface morphology with excellent residual stresses.

2. Experimental procedure

2.1. Materials and method

The steel substrate having diameter 8 mm was grounded to get uniform surface, using 120-2000 grit emery paper and after that polished with diamond paste. Table 1 shows the chemical composition substrate material. High quality Mo composite coating were deposited on the substrate using atmospheric HVOF coating process. Initially carbon powder, Cr_3C_2, Mo, and NiCr powder were blended together in the mixer, to get composite powder for deposition. A HVOF torch, PT F4 was used for deposition of composite powders. Deposition rate of composite powder was 70±6 g/min. The Mo based composite coating was having thickness ranged from 350-450µm. The deposition parameters to developed Mo based composite coating are summarized in Table 2. The FESEM was used to characterize the

Accepted Manuscript

topography of coated surface, while EDS was employed to examine the components such as Fe (iron), Mo (molybdenum), C (carbon), and Cr (chromium) peak present in the coating.

Table 1 chemical composition of substrate material

Element	C	S	Cr	Si	P	Mo	Ni	Cu	Mn
% Composition	0.16	0.01	0.10	0.24	0.01	0.02	0.02	0.17	1.07

Table 2 Experimental deposition conditions of Mo composite coating

Sr. no.	**Process Parameters**	**Values**
1.	Powder driving Temperature (^oC)	120
2.	Deposition Process	HVOF
3.	Hydrogen Flow Rate (mm^3/min.)	13
4.	Argon Flow Rate (mm^3/min.)	112
5.	Mixing powder time (min.)	90
6.	Gun Angle during spray (degree)	30
7.	Gun Feed (mm/min.)	10
8.	Cooling air pressure (Pa)	47

2.2. Percent (%) change in residual stress measurement

The FESEM-EDS of coated samples was carried out on JEOL made FESEM with EDS. The integrated EDS system was used to evaluate the presence of carbon (C) peak on the coated surface. HRXRD was carried out using PANalytical X'Pert system and was used to confirm the elemental composition present in the composite coating.

Residual stress test war used to compute the adhesion, quality, and performance of the developed Mo composite coating. The residual stresses were measured using of μ-X360, XRD residual stress analyzer based on cos(α) method, having CCD camera support, LED marker, and sensor unit.

Table 3 and Table 4 shows the design parameters and experimental results of Mo composite coating for Taguchi L_9 experiments respectively of % change in residual stress.

Accepted Manuscript

Table 3 Design parameters of Mo composite coating for Taguchi L_9

Process Parameter	Test 1	Test 2	Test 3	Designation
Frequency (Hz)	7	14	21	A
Temperature(^{o}C)	110	220	330	B
Load (N)	15	30	45	C

Table 4 Taguchi L_9 OA experimental results

Experiments	Frequency (Hz)	Temperature (^{o}C)	Load (N)	Percentage (%) change in residual stress			S/N ratio (db)
				R_1	R_2	R_3	
1	7	110	15	40.11	40.11	40.79	-32.11
2	7	220	30	30.91	30.84	30.65	-29.77
3	7	330	45	34.11	32.85	32.92	-30.44
4	14	110	30	42.12	40.19	40.21	-32.22
5	14	220	45	37.11	35.85	38.92	-31.43
6	14	330	15	23.56	24.35	25.54	-27.78
7	21	110	45	30.25	23.52	21.65	-28.10
8	21	220	15	15.85	18.61	15.23	-24.41
9	21	330	30	13.58	13.47	14.08	-22.74
T= overall mean of % change in residual stress = 29.16							

3. Results and discussions

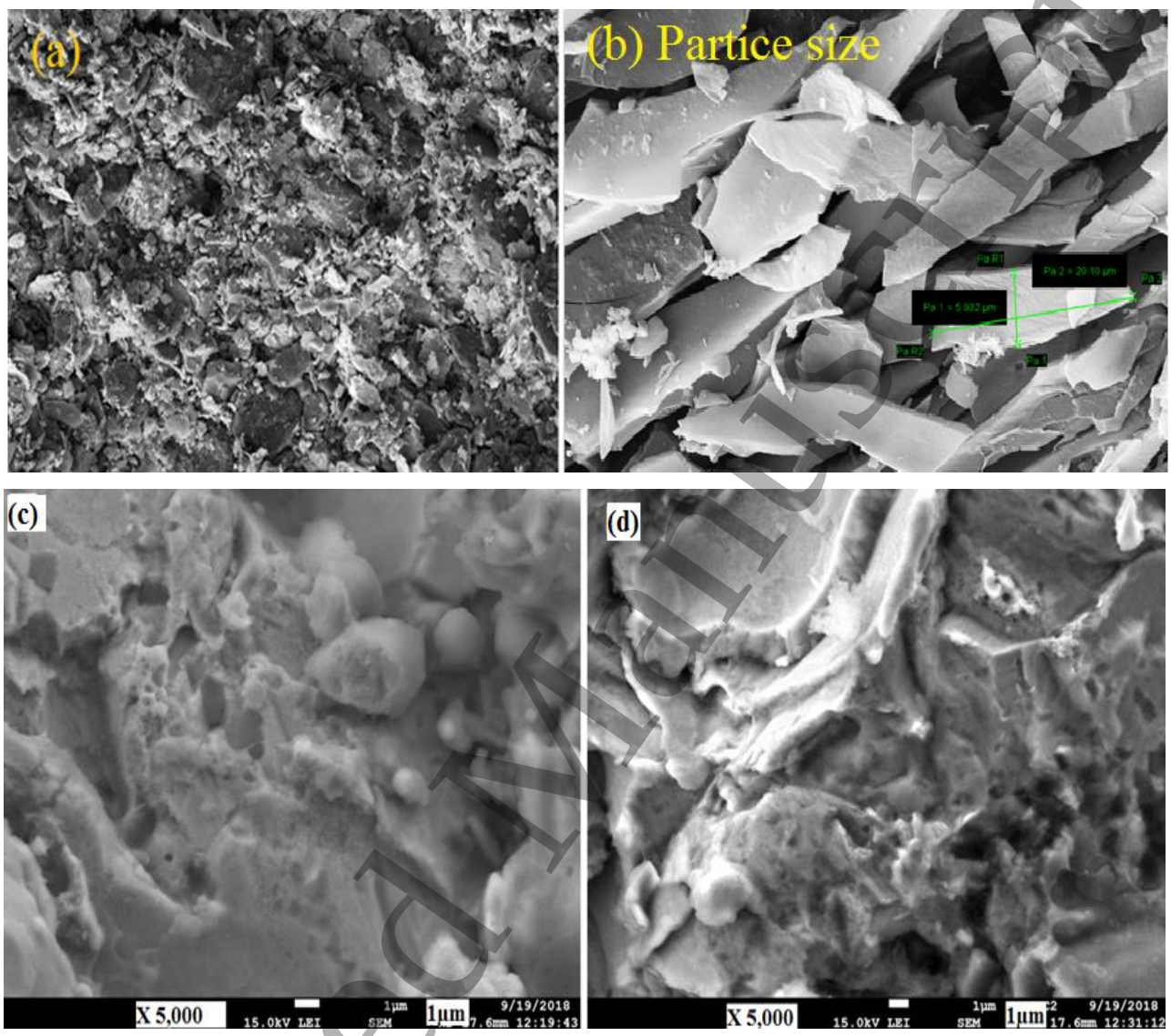
3.1.Coating Characterization

The FESEM micrographs and EDS results of composite coating samples are presented in fig. 1(a-e) which shows a desirable thermal and HVOF coating. The thickness of Mo composite coatings was ranged from 350-450±5µm and also having a surface roughness of 0.1±0.05µm. It is also noted from the micrographs that a curved laminar structure is formed with oxide layer in between as the oxide layer formation is confirmed due to the presence of O (oxygen) in EDS result. In addition to uniform coating, some voids and cracks are noticed, which may be attributed due formation of oxides during polishing process. The EDS analysis was done in order to confirm the element present in the composite coating. The EDS result shows the presence of Mo, Chromium, Nickel, Iron, Carbon, and Sulphur elements which confirms the formation of carbides, sulphides and oxides in the Mo composite coating.

HRXRD patterns shows the 2theta angle peak was at 40.4^o and 73.6^o, which confirms the presence of Cr_2C_2 and $Cr_{0.5}Mo_{0.5}$ respectively as shown in fig 1(e). Molybdenum carbide and oxide peaks were identified in the composite coating. This peak augmentation is recognized to the following; (i) Dislocation density increased with stacking faults, (ii) reduction in

Accepted Manuscript

crystallite size (iii) formation of low angle grain boundary. The presence of hard structured carbide and oxide layer formed on the steel substrate, shows excellent mechanical properties (residual stress).



Accepted Manuscript

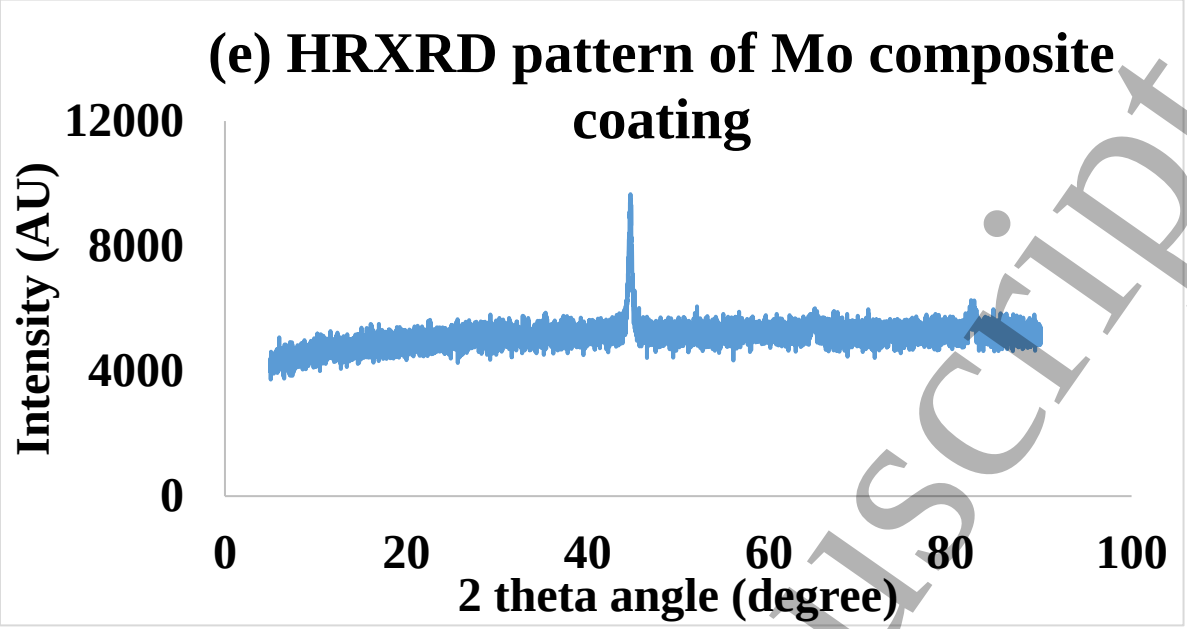


Fig. 1 (a-d) FESEM (e) HRXRD pattern of Mo composite coating

3.2.Effects of parameters on % change in residual stress

Table 5 shows the calculation of S/N ratio and main effects for all process parameters. The deviation in % change in residual stresses in plotted in fig. 2.

Table 5 Average values and main effects of residual stress

Parameter	Trial	Frequency (Hz)		Temperature (°C)		Load (N)	
Data		Raw Data	S/N Ratio	Raw Data	S/N Ratio	Raw Data	S/N Ratio
Average values (weight loss)	T1	35.43	-30.81	34.81	-30.77	27.12	-28.10
	T2	28.21	-28.54	34.20	-30.48	28.45	-28.24
	T3	23.82	-26.99	18.47	-25.05	31.90	-29.99
Main effect of weight loss	T2-T1	-7.22	2.27	-0.60	0.297	1.32	-0.141
	T3-T2	-4.39	1.55	-15.73	5.39	3.45	-1.75
Difference (T3-T2)-(T2-T1)		2.83	-0.72		5.098	2.13	-1.60

Fig. 2 (a) shows variation in % change in residual stress with frequency. It was observed from experimental results that maximum residual stress is attained at 7 Hz frequency while minimum was obtained at 21 Hz frequency. It was also observed that S/N ratio was maximum

Accepted Manuscript

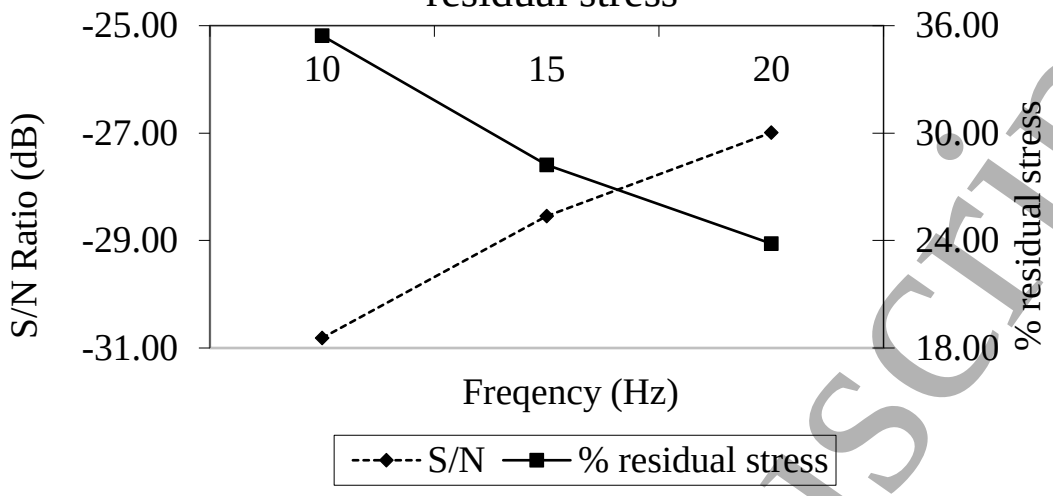
at 21 Hz frequency and minimum at 7 Hz frequency. The experiments conducted by Tyagi et al. [15] shows that as the sliding velocity increases, residual stresses decreases due to hard structured coating, graphitization effect and lower thermal expansion coefficient [15-17].

It was observed from the deviation of temperature with respect to % change in residual stress is shown in fig. 2(b). Experimental result obtained for composite coating shows that the % change in residual stresses is minimum at 330^{o}C temperature while maximum at 110^{o}C temperature. On the other hand, the S/N ratio is minimum at 110^{o}C temperature, while maximum at 330^{o}C temperature. The reduction in residual stress values is accredited due to lower thermal expansion coefficient results in lower thermal residual stresses [15-17]. The residual stress increases and tensile stress was generated on coated samples, as the temperature increased from 30-350^{o}C. This % change in residual stresses was accredited due to increase in operational temperature upto 400^{o}C or due to the formation of micro-structured layer which act as a deformable layer and also absorb the residual stresses developed on the coating as compared to Nano-structured ones as discussed by Skordaris *et al*. [16], in his research.

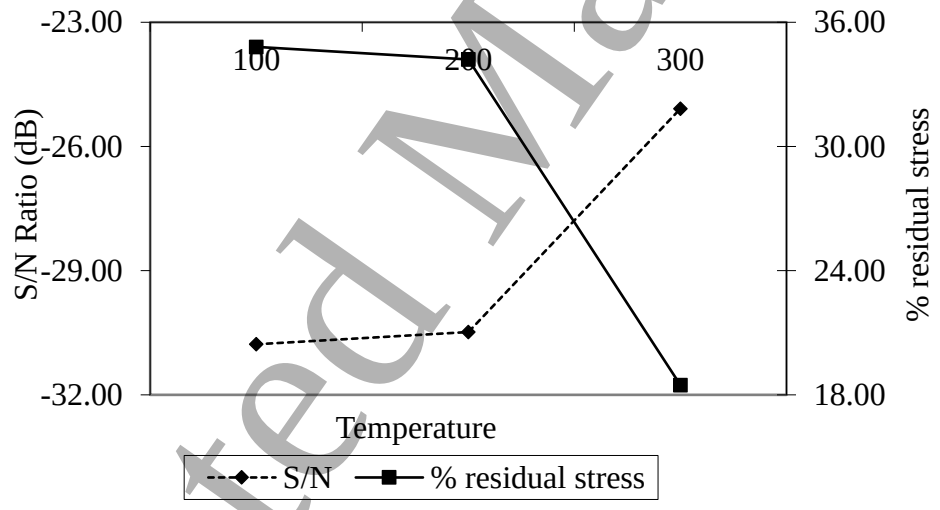
The deviation in load with respect to % change in residual stress of Mo based composite coating in shown in fig. 2(c). Experimental results that the maximum residual stress is attained at 45 N load while minimum was obtained at 15 N load. It was also observed that the S/N ratio is maximum at 15 N load and minimum at 45 N load. The experiments conducted by Tyagi et al. [15] shows that as the load increases, the % change in residual stresses also increases due to development in tensile stress, and lower thermal expansion coefficient [15-17].

Accepted Manuscript

(a) Effect of Sliding velocity on S/N and % residual stress



(b) Effect of Temperature on S/N and % residual stress



Accepted Manuscript

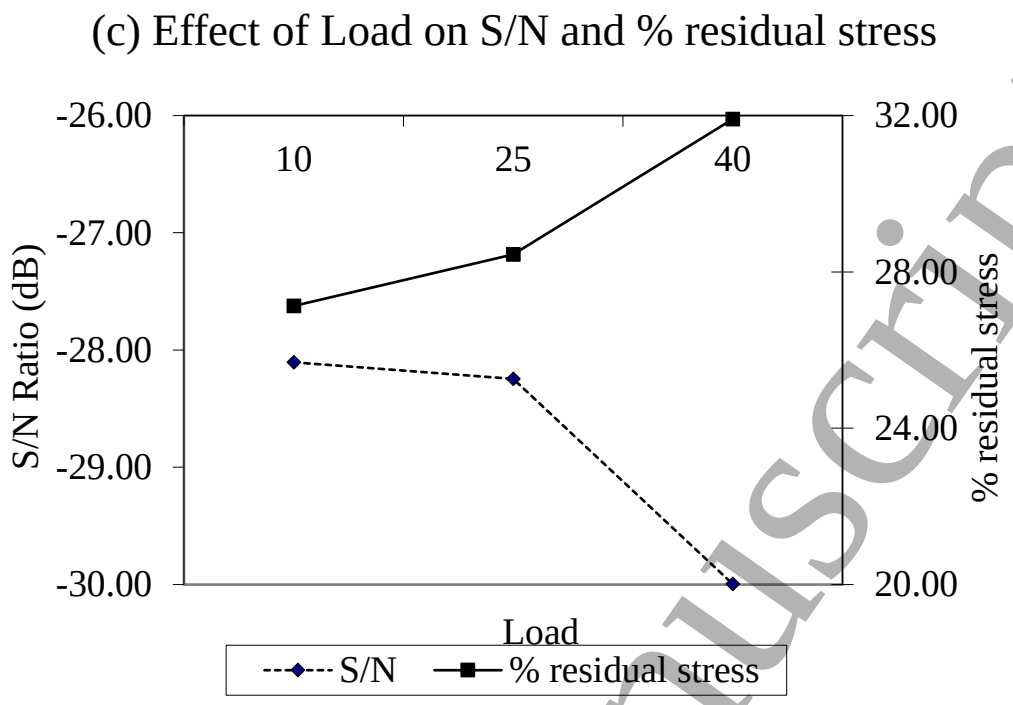


Fig. 2 shows deviation of parameters (a) Frequency (Hz) (b) Temperature (ºC) (c) Load (N) Vs % change in residual stress

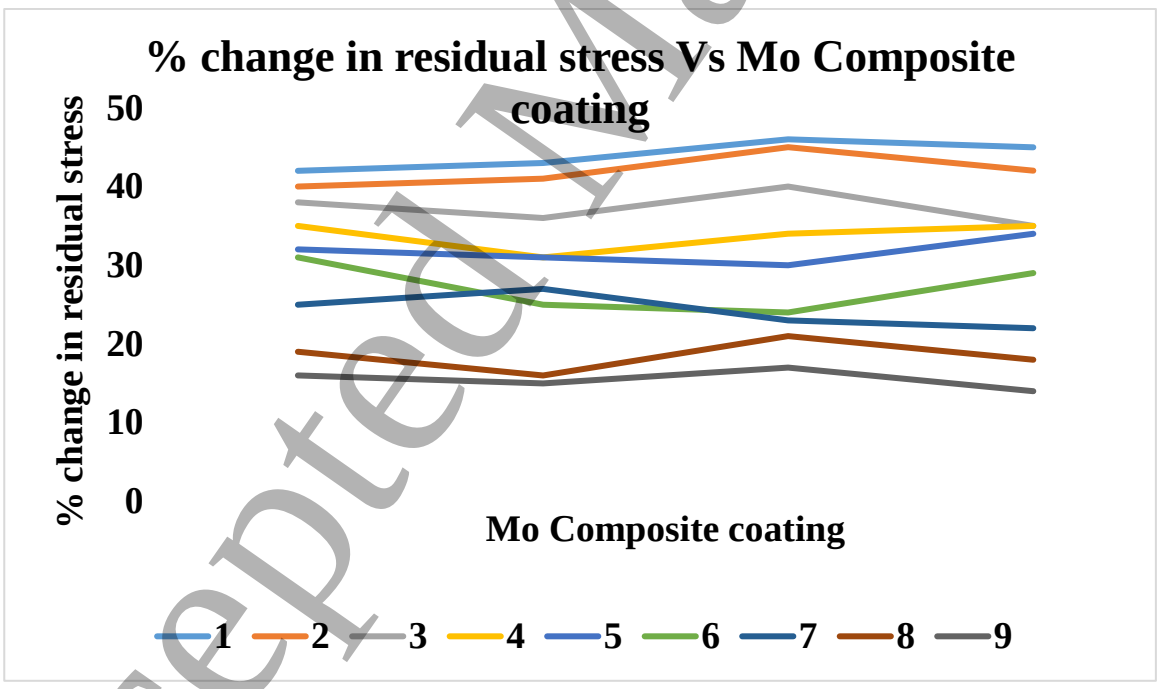


Fig. 3 shows variation in Mo composite coating Vs % change in residual stress

For better results, lower-the-better type characteristic was selected. Fig. 3 shows variation of residual stress verses Mo composite coating. From fig. 2, it was quite evident from the experimental results that the optimum and minimum values is at test 3 of parameter A (21Hz), and B (300ºC) and test 1 of the parameter C (15N) of % change in residual stress.

Accepted Manuscript

The XRD residual stress is a biaxial, non-destructive residual stress measurement technique. This XRD technique is versatile and can be used quantitatively to sense micro and macro residual stress separately. Fig. 4(a-c) shows variation of Debye ring and Distortion of coated samples. The Debye ring formed by a single short duration X-ray exposure of incidence ray represents the residual stress concentration in the substrate material. The concentration of residual stress are represented by different color combinations such as red color shows maximum and blue color represent minimum concentration. The experiments conducted by Tyagi et al. [15] and Skordaris et al. [16] shows that as the load increases, the residual stresses also increases due to development in tensile stress, and lower thermal expansion coefficient [15-17]. Biberger et al. [18] investigated thermal spray coating with hard chrome coating and observed that as the temperature of piston ring was increased from 150 to 200°C, wear volume decreases from $2.49*10^{-3}$ to $1.81*10^{-3}$ mm^3. The residual stress increases and tensile stress was generated on coated samples, as the temperature increased from 30-350^{o}C. This % change in residual stresses was accredited due to increase in operational temperature upto 400^{o}C or due to the formation of micro-structured layer which act as a deformable layer and also absorb the residual stresses developed on the coating as compared to Nano-structured ones as discussed by G. Skordaris *et al*. [16], in his research.

Accepted Manuscript

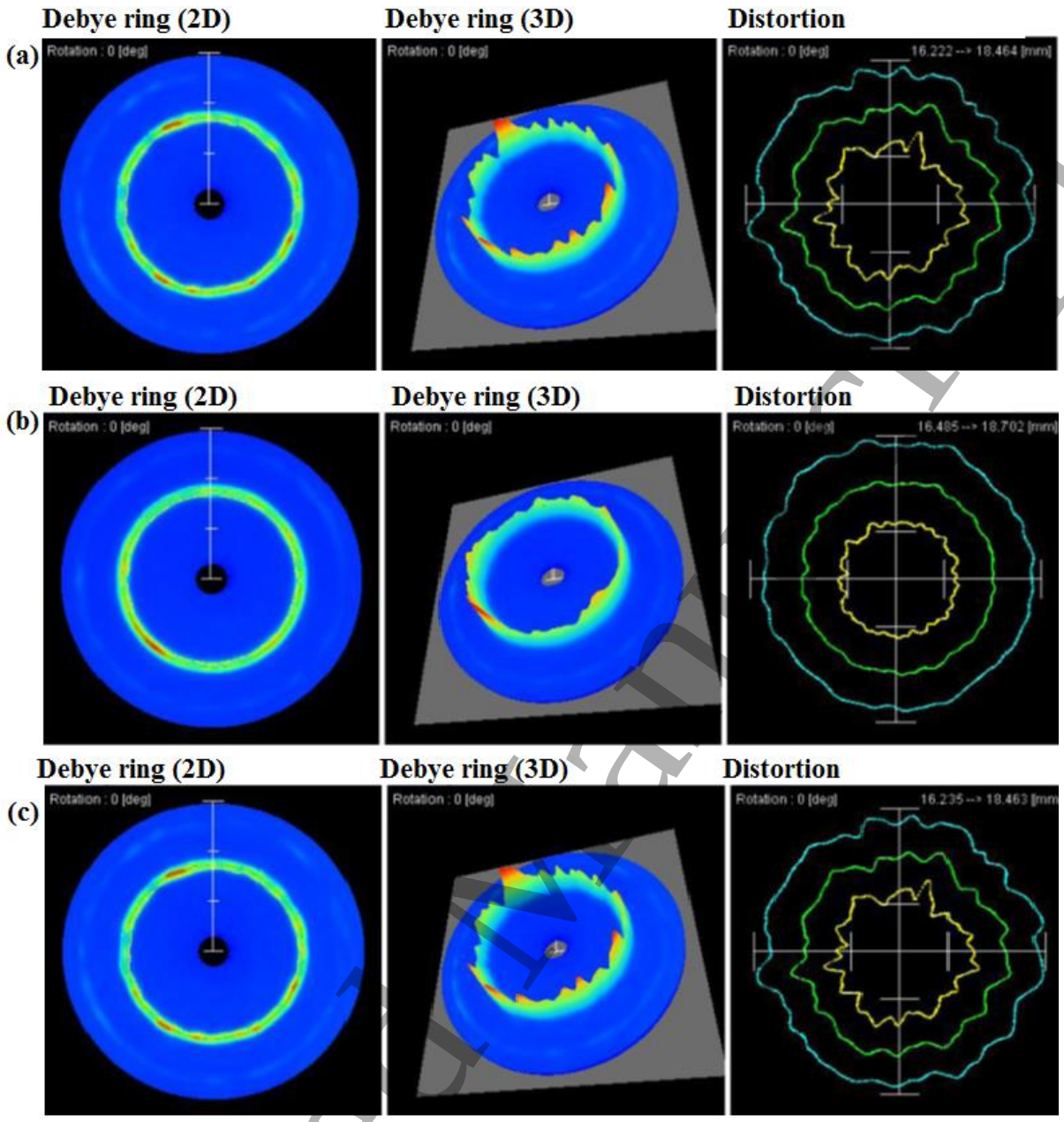


Fig. 4(a-c) shows variation of Debye rings and Distortion of coated samples

ANOVA method was used to analyze and identify the significance of process parameters of Mo composite coating. Table 6 shows the deviation of pooled ANOVA. It was quite evident from the experimental results that parameters A, B and C substantially effects residual stress of Mo based composite coating. Experimental results also concluded that for % change in residual stress, temperature (64.19 %) has maximum % contribution followed by frequency (21.82 %) and load (6.31 %).

Accepted Manuscript

Table 6 Pooled ANOVA values of residual stress

Source type	SS		DOF		V		F-Ratio		SS'		P (%)	
	Raw data	S/N data	Raw data	S/N data	Raw data	S/N data	Raw data	S/N data	Raw data	S/N data	Raw data	S/N data
Frequency (Hz)	618.57	22.16	2	2	309.28	11.08	46.43	64.25	605.25	21.82	25.70	24.42
Temperature (°C)	1544.69	61.58	2	2	772.34	30.79	114.96	178.56	1531.37	61.23	64.19	67.85
Load (N)	109.71	6.65	2	2	54.85	3..28	8.23	19.29	96.39	6.31	4.56	7.33
E (Pooled)	133.20	0.34	20	2	6.66	0.17	-	-	173.16	1.38	5.53	0.38
Total	2403.18	90.75	26	8	-	-	-	-	2406.18	90.75	100	100

3.2.1. Optimal characteristics estimation

3.2.1.1.Response characteristics for optimal values

Table 7 shows the significant parameters mean values at optimum level of % change in residual stress. Equation 1 [19-23] can be used to calculate the optimum value for residual stress.

$$\mu\%\ \text{change in residualstress} = A3 + B3 + C3 - 2T \quad (1)$$

$$= 11.09$$

Equations 2 and 3 was used to calculate the CI_{POP} and CI_{CE} values for 95 % confidence interval.

The ANOVA values attained is:

$$\text{CIce} = \sqrt{Fa(1, fa)Ve[\frac{1}{Neff}, \frac{1}{R}]} \quad (2)$$

$CI_{CE} = \pm 3.71$

$$\text{CIpop} = \frac{\sqrt{\text{Fa(1,fe)Ve}}}{\text{Neff.}} \quad (3)$$

$CI_{POP} = \pm 2.45$

R= 3; V_e = 6.66; N= 27; $N_{eff.}$ = 3.86; f_a = 20; $F_{0.05}$ (1, 20) = 3.49

Table 7 mean values at optimum level of % change in residual stress

Tests	% change in residual stress
A_3	23.82
B_3	18.47
C_1	27.12
T	29.16

The predicted optimum values for % residual stress is given by:

Accepted Manuscript

CI_{CE} : 7.38 < μ < 14.8

CI_{POP} : 8.64 < μ < 13.54

The predicted optimal values at the selected test of parameters are:

Test 3 of frequency (A_3) = 21 Hz

Test 3 of temperature (B_3) = 330°C

Test 1 of load (C_1) =15N

3.2.1.2.Confirmation experiments

For calculation and estimation of optimal situation, three confirmation tests were performed for process parameters. The process parameters are set, at test 3 of frequency (A_3) at 21 Hz, test 3 of temperature (B_3) at 330°C, test 1 of load (C_1) at15N. For Mo composite coating, the average value of residual stress was measured as 11.09 which is in between the optimal range.

4. Conclusions

The aim of present work was principally focus upon development of Mo composite coating using HVOF thermal spray coating process and to enhance residual stresses using Taguchi approach.

- Composite coating shows presence of Cr_3C_2, MoO_2 and Mo_2C along with Cr_2C_2, hard phased which was confirmed by HRXRD and FESEM results.
- Experimental results also concluded that for % change in residual stress, temperature (64.19 %) has maximum % contribution followed by frequency (21.82 %) and load (6.31 %).
- For 95% confidence interval, the predicted optimal range is CI_{POP}: 8.64 < μ < 13.54
- The predicted mean range of is CI_{CE}: 7.38 < μ < 14.8.

References

1. Directive 1999/45/EC of the European Parliament and of the Council. Directive related to the classification, packaging and labelling of dangerous preparations. Application to Greases and Ingredients. Official Journal of the European Communities, L 200/1, July 30, 1999.
2. Oil Pollution Prevention and Response; Non-Transportation-Related Facilities, Part III; Final Rule. 30.06.2000, 40 CFR Part 112. Federal Register Vol. 65, No. 127, Rules and Regulations.
3. Shahmohamadi, H., Mohammadpour, M., Rahmani, R., Rahnejat, H., Garner, C. P., & Howell-Smith, S. (2015). On the boundary conditions in multi-phase flow through the piston ring-cylinder liner conjunction. Tribology International,90, 164-174.

Accepted Manuscript

4. Kapsiz M, Durat M, Ficici F (2011) Friction and wear studies between cylinder liner and piston ring pair using Taguchi design method. Adv Eng Softw, 2011; 42; 595–603.
5. Yan J, He Z, Wang Y, et al. Microstructure and wear resistance of plasma-sprayed molybdenum coating reinforced by MoSi2 particles. J Therm Spray Technol. 2016; 25; 1322–1329.
6. Matikainena V, Bolelli G, Koivuluoto H, et al. Sliding wear behaviour of HVOF and HVAF sprayed Cr3C2-based coatings. Wear. 2017. doi:10.1016/j.wear.2017.04.001
7. Zhou W, Zhou K, Deng C, et al. Hot corrosion behavior of HVOF-sprayed Cr3C2-WC-NiCoCrMo coating. Ceram Int. 2017. doi:10.1016/j.ceramint.2017.04.109
8. Shailesh Mani Pandey, Qasim Murtaza, R. S. Walia, Study of dry wear behavior and morphological characteristic of 60%Mo-20%NiCr-10%CrC-10%Mo+Fe based alloy coating by atmospheric plasmaspray technique, Advances in Materials and Processing Technologies, DOI: 10.1080/2374068X.2017.1334997.
9. Ankit Tyagi, R S Walia, Qasim Murtaza, "Tribological behavior of HVOF carbon coating for wear resistance applications" Mater. Res. Express (Accepted 2019 https://doi.org/10.1088/2053-1591/ab555d).
10. Hong S, Wu Y, Wang B, et al. The effect of temperature on the dry sliding wear behavior of HVOF sprayed nanostructured WC-CoCr coatings. Ceram Int. 2017; 43; 458–462.
11. Ankit Tyagi, R.S. Walia, Qasim Murtaza, Shailesh Mani Pandey, Pawan Tyagi, Bharat Bajaj, A critical review of Diamond like Carbon Coating for Wear Resistance Applications, International Journal of Refractory Metals & Hard Materials, 2019; 78; 107–122.
12. Ankit Tyagi, S M Pandey, Kalpna Gupta, R.S. Walia, Qasim Murtaza, Kumar Krishen, "Tribological behavior of sustainable Carbon based Composite coating for wear resistance applications" Mater. Res. Express (Accepted 2019 https://doi.org/10.1088/2053-1591/ab52b4).
13. Ankit Tyagi, SM Pandey, Qasim Murtaza, R.S. Walia, Mohit Tyagi, "Tribological behavior of carbon coating for piston ring applications using Taguchi approach", Materials today proceeding (Accepted 2019).
14. V. Kavimani *, K. Soorya Prakash, "Tribological behaviour predictions of r-GO reinforced Mg composite using ANN coupled Taguchi approach", Journal of Physics and Chemistry of Solids 110 (2017) 409–419.

Accepted Manuscript

15. Tyagi A., Walia, RS., Murtaza Q., Tribological behavior of temperature dependent environment friendly thermal CVD diamond coating, Diamonds and Related Materials, 96 (2019) 148–159.
16. G. Skordaris, K.-D. Bouzakis, P. Charalampous, T. Kotsanis, E. Bouzakis, O. Lemmer, Effect of structure and residual stresses of diamond coated cemented carbide tools on the film adhesion and developed wear mechanisms in milling, CIRP Annals - Manufacturing Technology, 2016; 65; 101–104.
17. Bouzakis K-D, Skordaris G, Bouzakis E, Charalampous P, Kotsanis T, Tasoulas D, Kombogiannis S, Lemmer O, Effect of the Interface Fatigue Strength of NCD Coated Hardmetal Inserts on Their Cutting Performance in Milling, Diamond and Related Materials, 2015; 59; 80–89.
18. J.A. Picas, A. Forn, G. Matthaus, HVOF coatings as an alternative to hard chrome for pistons and valves, Wear 261 (2006) 477–484
19. Tyagi A., Murtaza Q., Walia, RS., Pandey SM., Characterization and parametric optimization of tribological properties of Mo blend composite coating, Material Research Express, (Accepted, 2019).
20. T Goyal, RS Walia, TS Sidhu, Study of coating thickness of cold spray process using Taguchi method, Materials and Manufacturing Processes 27 (2), 185-192
21. RS Walia, HS Shan, P Kumar, Parametric optimization of centrifugal force-assisted abrasive flow machining (CFAAFM) by the Taguchi method, Materials and Manufacturing Processes 21 (4), 375-382
22. BS Brar, RS Walia, VP Singh, Electrochemical-aided abrasive flow machining (ECA^2FM) process: a hybrid machining process, The International Journal of Advanced Manufacturing Technology 79 (1-4), 329-342
23. T Goyal, RS Walia, TS Sidhu, Effect of parameters on coating density for cold spray process, Materials and Manufacturing Processes 27 (2), 193-200

Accepted Manuscript

arXiv:1912.10237v1 [q-fin.MF] 21 Dec 2019

Comparative Study of Two Extensions of Heston Stochastic Volatility Model

Gifty Malhotra
giftymalhotradtu@gmail.com

R. Srivastava
rsrivastava@dce.ac.in

H.C. Taneja
hctaneja@dce.ac.in

Department of Applied Mathematics, Delhi Technological University,
Delhi(India)-110042

Abstract

In the option valuation literature, the shortcomings of one factor stochastic volatility models have traditionally been addressed by adding jumps to the stock price process. An alternate approach in the context of option pricing and calibration of implied volatility is the addition of a few other factors to the volatility process. This paper contemplates two extensions of the Heston stochastic volatility model. Out of which, one considers the addition of jumps to the stock price process (a stochastic volatility jump diffusion model) and another considers an additional stochastic volatility factor varying at a different time scale (a multiscale stochastic volatility model). An empirical analysis is carried out on the market data of options with different strike prices and maturities, to compare the pricing performance of these models and to capture their implied volatility fit. The unknown parameters of these models are calibrated using the non-linear least square optimization. It has been found that the multiscale stochastic volatility model performs better than the Heston stochastic volatility model and the stochastic volatility jump diffusion model for the data set under consideration.

Keywords. Stochastic volatility; Multiscale stochastic volatility; Mean reversion; Option pricing; Time scales; Jump diffusion

1 Introduction

The derivative pricing model proposed by Black and Scholes[5] assumes the volatility to be constant and asset log-return distribution as Gaussian. Empirically, the volatility is not constant but it smiles and the log-return distributions are non-Gaussian in nature characterised by heavy tails and high peaks. A wide range of research has been done to improve upon classical Black-Scholes model. The model has been extended to include either constant volatility with jumps(e.g. jump diffusion (JD) models of Merton[15]

and Kou[13]) or to consider volatility itself as a continuous time stochastic process(e.g. stochastic volatility models given by Hull and White[11], Scott[17], Wiggins[21], Stein and Stein[20], Heston[10] and Ball and Roma[3], etc.). Stochastic volatility models allow the volatility to fluctuate randomly and are able to incorporate many stylized facts about volatility namely volatility smile and skew, mean reversion and leverage to name a few.

In the single factor stochastic volatility models, Heston stochastic volatility model is most popular as it gives a fast and easily implemented semi closed form solution for the European options and is relatively economical from the computational point of view. Despite its success and popularity, it has some shortcomings. The model is unable to fit implied volatility across all strikes and maturities particularity for the options with short expiry [12]. Also, Shu and Zhang[19] obtained that the Heston model overprices out-of-money (OTM) and short-term options and it underprices in-the-money (ITM) options.

In the option valuation literature, the shortcomings of one factor stochastic volatility models have traditionally been addressed by adding jumps to the stock price process(e.g. stochastic volatility jump diffusion (SVJ) models of Bates[4], Scott[18] and Pan[16], etc.). Jumps are added to the stock price dynamics of a stochastic volatility model which improve its pricing performance for the short-term options [2]. An alternate approach is the consideration of multiscale stochastic volatility (MSV) models to address the shortcomings of one factor stochastic volatility models (see [9, 14]). In these models, volatility is driven by several factors varying at different time scales. Alizadeh et al.[1] found the evidence of two factors of volatility with one highly persistent factor and other quickly mean reverting factor. Extending this idea, Fouque et al.[8] proposed a two factor stochastic volatility model with one fast mean reverting factor and another slowly varying factor. Christoffersen et al.[6] empirically showed that the two-factor models improve one factor models in the term structure dimension as well as in the moneyness dimension.

As both type of models (SVJ or MSV) are the extensions of classical single factor stochastic volatility models, this motivated us to study and compare these two approaches in context of Heston stochastic volatility model. For this, we have considered two extensions of Heston stochastic volatility model. One is the stochastic volatility jump diffusion model proposed by Yan and Hanson[22] which is an extension of Heston stochastic volatility model by adding jumps to the stock price process with log-uniformly distributed jump

amplitude. The another model is the multiscale stochastic volatility model proposed by Fouque and Lorig[7], in which a fast mean-reverting factor is additionally considered in the framework of Heston stochastic volatility model. These two models are compared with each other, and also with the Heston stochastic volatility model using S&P 500 index options data. Firstly, the model parameters are calibrated using non-linear least square optimization. Then the models' fit to the market implied volatility is captured against log moneyness at different time to maturity. The mean relative error of models' prices with market data is also calculated. We have obtained that the multiscale stochastic volatility model performs better than the other two models.

The rest of the paper is organised as follows: The underlying models has been explained in Section 2. The empirical analysis has been conducted in Section 3, where the calibration of the models' parameters, models' fit to market implied volatility and mean relative error of model prices with market data has been reported and the results obtained are discussed. The conclusion has been given in Section 4.

2 Models Under Consideration

Firstly, the two models to be considered for the empirical analysis has been explained.

2.1 Stochastic Volatility Jump Diffusion Model

Yan and Hanson [22] proposed a SVJ model which considers the log-uniform distribution of the jump amplitudes in the stock price process. The model is explained below:

Let X_t be the stock price at time t whose dynamics under the risk-neutral probability measure P^* is

$$dX_t = X_t((r - \lambda\overline{J})dt + \sqrt{V_t}dW_t^x + J(U)dN_t) \tag{1}$$

where r is the risk free interest rate and $J(U)$ is the Poisson jump-amplitude with mean $\overline{J}$. The variance V_t follows the CIR process given by

$$dV_t = \kappa(\theta - V_t)dt + \sigma\sqrt{V_t}dW_t^v \tag{2}$$

with κ as the rate of mean-reversion, θ as the long-run mean value and σ as the volatility of variance. The condition $2\kappa\theta \geq \sigma^2$ must be satisfied to ensure the positivity of the process (2). W_t^x and W_t^v are the standard Brownian motions for the stock price process and the volatility process respectively with correlation

$$E[dW_t^x.dW_t^v] = \rho_{xv}dt$$

U is the amplitude mark process which is assumed to be uniformly distributed with density

$$\varphi_U(u) = \begin{cases} \frac{1}{n-m} & \text{if } m \leq u \leq n, \\ 0 & \text{otherwise} \end{cases}$$

and is given by

$$U = \ln(J(U) + 1)$$

N_t is the standard Poisson jump counting process with jump intensity λ, $J(U)dN_t$ is the Poisson sum which is given as

$$J(U)dN_t = \sum_{i=1}^{dN_t} J(U_i)$$

here U_i is the ith jump-amplitude random variable and $\overline{J}$, the mean of jump-amplitude J, is given as

$$\overline{J} = E[J(U)] = \frac{e^n - e^m}{n - m} - 1.$$

Under this model, the pricing formula for the European call option, in terms of log stock price $s = \ln(x)$, is given as:

$$C_{svj} = e^s P_1(s, v, t, K, T) - Ke^{-r(T-t)} P_2(s, v, t, K, T) \tag{3}$$

where $v = V_t$ is the variance at time t, T is the maturity time, K is the strike price and r is the risk free interest rate. The subscript svj in the price C_{svj} is just to specify the price obtained from SVJ model. The same convention is also followed for Heston and MSV model.
For $j = 1, 2$,

$$P_j(s, v, t, K, T) = \frac{1}{2} + \frac{1}{\pi} \int_0^\infty Re\left[\frac{e^{-i\phi \ln K} f_j(s, v, t, \phi, T)}{i\phi}\right] d\phi \tag{4}$$

where the characteristic function f_j of P_j is

$$f_j(s, v, t, \phi, t+\tau) = e^{A_{(1j)}(\tau,\phi)+A_{(2j)}(\tau,\phi)v+i\phi s+\beta_j(\tau)} \tag{5}$$

with $\tau = T - t$ and $\beta_j(\tau) = r\tau\delta_{j,2}$; $\delta_{j,2} = 1$ for $j = 2$ and 0 for $j = 1$. The other terms are

$$A_{(1j)}(\tau,\phi) = r\phi i\tau - (\lambda\overline{J}i\phi + \lambda\overline{J}\delta_{j,1} + r\delta_{j,2})\tau + \lambda\tau\left[\frac{e^{(i\phi+\delta_{j,1})n} - e^{(i\phi+\delta_{j,1})m}}{(n-m)(i\phi+\delta_{j,1})} - 1\right] + A'_{(1j)}(\tau,\phi) \tag{6}$$

$$A_{(2j)}(\tau,\phi) = \frac{b_j - \rho\sigma\phi i + d_j}{\sigma^2}\left(\frac{1-e^{d_j\tau}}{1-g_j e^{d_j\tau}}\right) \tag{7}$$

and

$$A'_{(1j)}(\tau,\phi) = \frac{\kappa\theta}{\sigma^2}\left[(b_j - \rho\sigma\phi i + d_j)\tau - 2\ln\left(\frac{1-g_j e^{d_j\tau}}{1-g_j}\right)\right] \tag{8}$$

with

$$g_j = \frac{b_j - \rho\sigma\phi i + d_j}{b_j - \rho\sigma\phi i - d_j}$$

$$d_j = \sqrt{(\rho\sigma\phi i - b_j)^2 - \sigma^2(2\alpha_j\phi i - \phi^2)}$$

and

$$\alpha_1 = \frac{1}{2},\ \alpha_2 = \frac{-1}{2},\ b_1 = \kappa - \rho\sigma,\ b_2 = \kappa \tag{9}$$

The unknown parameters of this model are $\kappa,\ \theta,\ \sigma,\ \rho,\ v,\ \lambda,\ m$ and n.

After the SVJ model, the MSV model of Fouque and Lorig [7] is given below.

2.2 Multiscale Stochastic Volatility Model

Fouque and Lorig [7] extended the Heston stochastic volatility model to a MSV model by considering an additional fast mean-reverting volatility factor in the Heston stochastic volatility model. This model is given below.

Under P^*, the dynamics of stock price X_t is given as

$$dX_t = rX_t dt + \eta_t X_t dW_t^x \tag{10}$$

here $\eta_t = \sqrt{V_t} f(Y_t)$. Y_t and V_t are respectively the fast and the slow scale factors of volatility with their dynamics given as

$$dY_t = \frac{V_t}{\epsilon}(m - Y_t)dt + \mu\sqrt{2}\sqrt{\frac{V_t}{\epsilon}}dW_t^y \tag{11}$$

and

$$dV_t = \kappa(\theta - V_t)dt + \sigma\sqrt{V_t}dW_t^v \tag{12}$$

W_t^x ,W_t^y and W_t^v are the standard Brownian motions for the stock price process and for the fast and the slow factors of volatility respectively with $E(dW_t^x.dW_t^y) = \rho_{xy}dt$, $E(dW_t^x.dW_t^v) = \rho_{xv}dt$, and $E(dW_t^y.dW_t^v) = \rho_{yv}dt$. ρ_{xy}, ρ_{xv} and ρ_{yv} are constants which satisfy $\rho_{xy}^2 < 1, \rho_{xv}^2 < 1, \rho_{yv}^2 < 1$ and $\rho_{xy}^2 + \rho_{xv}^2 + \rho_{yv}^2 - 2\rho_{xy}\rho_{xv}\rho_{yv} < 1$.

The fast factor of volatility, Y_t follows the OU process with the mean-reversion rate V_t/ϵ and volatility of volatility parameter $\mu\sqrt{2}\sqrt{\frac{V_t}{\epsilon}}$. $\epsilon > 0$ is very small so that Y_t is fast mean-reverting towards its long-run mean m. The slow volatility factor V_t, as already explained for SVJ model, is the square root process. It slowly reverts to its long-run mean θ.

Fouque and Lorig[7] used the perturbation technique to obtain the expression for European call option prices. The asymptotic expansion of price in powers of $\sqrt{\epsilon}$ is given as

$$C_{msv}^{\epsilon}(x, y, v, t) = C_0 + \sqrt{\epsilon}C_1 + \epsilon C_2 + ... \tag{13}$$

They obtained the first order approximation to the price of the European call option as

$$C_{msv}^{\epsilon}(x, v, t) \approx C_0(x, v, t) + \sqrt{\epsilon}C_1(x, v, t)$$

This price approximation is clearly independent of the fast factor of volatility and depends only on the slow volatility factor v. The approximated price is perturbed around the Heston price C_0 at the effective correlation $\rho_{xv} < f >$, where $< f >$ is the average of $f(y)$ with respect to long-run distribution of the volatility factor Y_t.

The first order approximation term C_1 is

$$C_1 = e^s Q_1(s, v, t, K, T) - Ke^{-r(T-t)}Q_2(s, v, t, K, T) \tag{14}$$

where $s = \ln x$. For $j = 1, 2$

$$Q_j(s, v, t, K, T) = \frac{1}{2} + \frac{1}{\pi}\int_0^{\infty} Re\left[\frac{e^{-i\phi \ln K} q_j(s, v, t, \phi, T)}{i\phi}\right]d\phi \tag{15}$$

The characteristic function q_j of Q_j is

$$q_j(s,v,t,\phi,t+\tau) = (\kappa\theta\hat{q_0}(\tau,\phi) + v\hat{q_1}(\tau,\phi))(e^{A'_{(1j)}(\tau,\phi)+A_{(2j)}(\tau,\phi)v+i\phi s}) \quad (16)$$

here

$$\hat{q_0}(\tau,\phi) = \int_0^\tau \hat{q_1}(z,\phi)dz,$$
$$\hat{q_1}(\tau,\phi) = \int_0^\tau B(z,\phi)e^{A_{(3j)}(\tau,\phi,z)}dz$$

with

$$A_{(3j)}(\tau,\phi,z) = (b_j - \rho\sigma\phi i + d_j)\frac{1-g_j}{d_j g_j}\ln\left(\frac{g_j e^{d_j\tau}-1}{g_j e^{d_j z}-1}\right)$$
$$B(\tau,\phi) = -(V_1 A_{(2j)}(\tau,\phi)(2\alpha_j\phi i - \phi^2) + V_2 A^2_{(2j)}(\tau,\phi)(\phi i) + V_3(2\alpha_j\phi^3 i + \phi^2)$$
$$+ V_4 A_{(2j)}(\tau,\phi)(-\phi^2)) \quad (17)$$

All the other terms are already given in Eq.(8) to Eq.(9). The unknown parameters of this model are $\kappa,\ \theta,\ \sigma,\ \rho,\ v,\ V_1,\ V_2,\ V_3$ and V_4.

In the next section, the empirical analysis is conducted to compare these models.

3 Empirical Analysis and Discussion of Results

For the empirical analysis, the data[1] of S&P 500 index options is considered from January 4, 2016 with maturity ranging from 30 days to 180 days and moneyness from 75% to 125%. The risk free rate of interest is the yield on 3-month U.S. government treasury bill. Firstly, the unknown parameters of each model are calibrated using non-linear least square optimization. Once the parameters are obtained, the models' fit to the market implied volatilities for the S&P 500 index are captured and plotted against log moneyness. To compare the pricing performance, the mean relative error of each model price is calculated corresponding to the market option price data. These methods are explained in following subsections.

[1]Data sharing is not applicable to this article as no new data were created or analyzed in this study.

3.1 Calibration of Model Parameters

The unknown parameters of Heston stochastic volatility model, SVJ model and MSV model are calibrated using the data of S&P 500 index options. Let M_1, M_2 and M_3 denote the parameter sets of unknown parameters of Heston, SVJ and MSV model respectively , such that

$$\begin{aligned} M_1 &= (\zeta, \rho, \sigma, \theta, v) \\ M_2 &= (\zeta, \rho, \sigma, \theta, v, \lambda, m, n) \\ M_2 &= (\zeta, \rho, \sigma, \theta, v, V_1, V_2, V_3, V_4) \end{aligned} \tag{18}$$

here all of these unknown parameters are already mentioned in Section 2 except ζ, which is obtained from the condition $2\kappa\theta \geq \sigma^2$ of the CIR process (2) such that $\zeta = \kappa - \frac{\sigma^2}{2\theta},\ \zeta \geq 0$. Thus, the rate of mean-reversion $\kappa = \zeta + \frac{\sigma^2}{2\theta}$ is obtained from the calibrated values of $\zeta,\ \sigma$ and θ.

These parameters are calibrated by non-linear least square optimization using *MATLAB2012b*. The objective function is defined as:

$$\Delta_h^2(M_1) = \sum_j \sum_{i(j)} (C_{mkt}(T_j, K_{i(j)}) - C_h(T_j, K_{i(j)}, M_1))^2 \tag{19}$$

$$\Delta_{svj}^2(M_2) = \sum_j \sum_{i(j)} (C_{mkt}(T_j, K_{i(j)}) - C_{svj}(T_j, K_{i(j)}, M_2))^2 \tag{20}$$

$$\Delta_{msv}^2(M_3) = \sum_j \sum_{i(j)} (C_{mkt}(T_j, K_{i(j)}) - C_{msv}(T_j, K_{i(j)}, M_3))^2 \tag{21}$$

where $C_{mkt}(T_j, K_{i(j)})$ is the market price of call option with maturity T_j. For each expiration T_j, the available collection of strike prices is $K_{i(j)}$. Similarly, for a particular value of T_j and $K_{i(j)}$, $C_h(T_j, K_{i(j)}, M_1)$, $C_{svj}(T_j, K_{i(j)}, M_2)$ and $C_{msv}(T_j, K_{i(j)}, M_3)$ are the prices of the European call options with expiration date T_j and exercise price $K_{i(j)}$, calculated from the Heston stochastic volatility model with parameter set M_1, SVJ model with the parameter set M_2 and MSV model with the parameter set M_3 respectively.

The optimal set of parameters M_1^*,M_2^* and M_3^* is obtained which satisfies

$$\begin{aligned} \Delta_h^2(M_1^*) &= min_{M_1}(\Delta_h^2(M_1)) \\ \Delta_{svj}^2(M_2^*) &= min_{M_2}(\Delta_{svj}^2(M_2)) \\ \Delta_{msv}^2(M_3^*) &= min_{M_3}(\Delta_{msv}^2(M_3)) \end{aligned} \tag{22}$$

Firstly, the optimal parameter set for the Heston stochastic volatility model, M_1^*, is calibrated. Once the M_1^* is obtained, the initial iteration for SVJ model is taken as $(M_1^*,\ 50,\ -0.01,\ 0.01)$ with the lower and upper bounds for the last three components as $(1,\ -1,\ 0)$ and $(100,\ 0,\ 1)$ respectively. Similarly, the initial iteration for MSV model is taken as $(M_1^*,\ 0.0001,\ 0,\ 0,\ 0)$ with the lower and upper bounds for last four components as $(-0.05,\ -0.05,\ -0.05,\ -0.05)$ and $(0.05,\ 0.05,\ 0.05,\ 0.05)$ respectively.

Using the optimal parameter set, the implied volatility fit for all the three models is obtained and is plotted against log moneyness $(log\frac{K}{X})$. The models fit are compared relative to market implied volatility (MV) data. It is given in Fig.1 to Fig.3 for time to maturity 30 days, 90 days and 180 days respectively.

The parameters are calibrated from the whole data but the results are given and discussed for the different maturity times, separately.

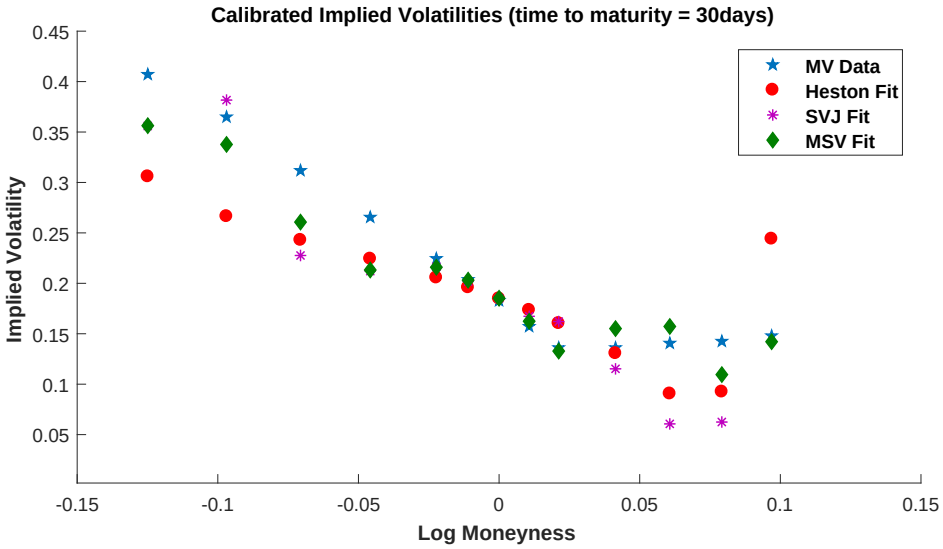


Figure 1: Models' fit to the implied volatilities of S&P 500 index with 30 days to maturity

Along with this, the mean relative error of the prices obtained from Heston stochastic volatility model, SVJ model and MSV model, with the

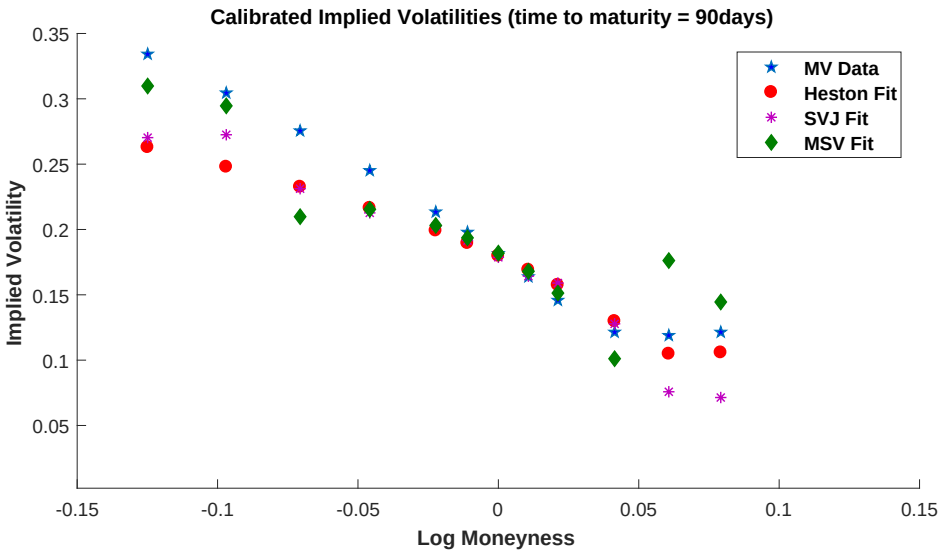


Figure 2: Models' fit to the implied volatilities of S&P 500 index with 90 days to maturity

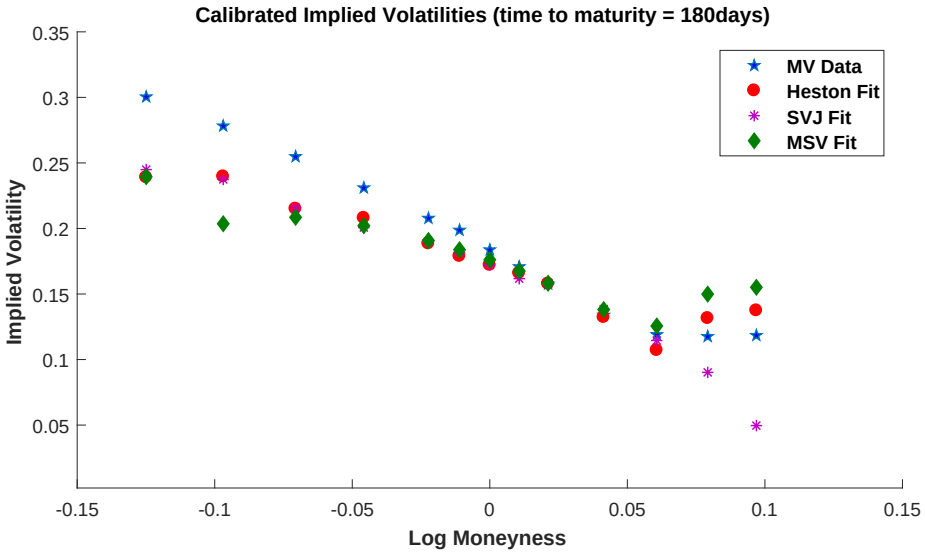


Figure 3: Models' fit to the implied volatilities of S&P 500 index with 180 days to maturity

market data of S&P 500 index option, is calculated at different maturity time. It is given in subsection 3.2.

3.2 Mean Relative Error

For a particular model with price $C_{model}(T_j, K_{i(j)}, \Theta)$ at different values of T_j and $K_{i(j)}$, the mean relative error (MRE) of model price with respect to market price, at time to maturity T_j is given as

$$MRE(j) = \frac{1}{N_j} \sum_{i(j)} \frac{|C_{model}(T_j, K_{i(j)}, \Theta) - C_{mkt}(T_j, K_{i(j)})|}{C_{mkt}(T_j, K_{i(j)})} \tag{23}$$

where N_j is the different number of call options that has expiry at time T_j, Θ is the optimal parameter set for the given model.

The mean relative error of Heston stochastic volatility model, SVJ model and MSV model is calculated for S&P 500 index data set. The maturity time is taken from 30 days to 180 days. Corresponding to a particular maturity, the strike prices range from 75% to 125%. The results are given in Table 1.

Table 1: The mean relative error of models prices with respect to market data.

	Models		
Maturity Time (T)	Heston	SVJ	MSV
30 days	0.0697	0.0499	0.0225
90 days	0.0874	0.0987	0.0456
180 days	0.0284	0.1070	0.0380

Now, we discuss the results obtained in Fig.1 to Fig.3 and in Table1. From the models fit to the implied volatility given in Fig.1 to Fig.3, it is clearly observable that the MSV model performs in an improved way in comparison to Heston stochastic volatility model and the SVJ model. For at-the-money (ATM) and near the money options, all the three models give equivalent results. The difference is observable for ITM and OTM options.

In Fig.1, the maturity time is short, that is 30 days. For such options, the Heston model fit to market implied volatility is not good. This supports the empirical findings that the Heston model poorly performs for short term

options. The SVJ model performs better than Heston model and MSV model for deep ITM options, but as the log moneyness value increases, MSV model outperforms both Heston model and SVJ model.

In Fig.2, the maturity time is medium, that is 90 days. The implied volatility fit of Heston model is improved for the OTM options. For ITM options, implied volatility fit of MSV model is better than the implied volatility fit of Heston model. The Heston model fit is equivalent to the SVJ model fit to market implied volatility.

In Fig.3, at the longer maturity, which is 180 days, all of the three models give almost similar fit for ITM options but for OTM options the Heston model outperforms the other two models. The implied volatility fit of MSV model is better than the fit of SVJ model to the market implied volatilities. Thus, out of SVJ model and MSV model, the overall fit of MSV model to the market implied volatility is better than SVJ model.

Additionally, from Table1, the pricing performance of three models is compared in terms of mean relative error of models prices with the market option price data. For the short and medium term options with maturity 30 and 90 days respectively, the mean relative error of MSV model is least. Thus the MSV model performs better than the SVJ model and Heston model in pricing. For maturity time 30 days, SVJ model performs better than Heston model in pricing, but for maturity 90 days, Heston model gives better pricing performance.

For the long term options with maturity 180 days, the MSV model performs better than SVJ model and Heston model outperforms the SVJ and MSV model.

Thus, out of SVJ model and MSV model, the overall pricing performance of MSV model is better than SVJ model for the data set under consideration.

4 Conclusion

The two extensions of Heston stochastic volatility model, already proposed in literature, are studied and compared in this paper on the basis of their fit to the market implied volatility and pricing performance. An empirical analysis is conducted on S&P 500 index options data and the results are obtained for all the three models. It has been obtained that for the data set under consideration, multiscale stochastic volatility performs better than the stochastic volatility jump model. Thus, the inclusion of additional volatility

factor to a stochastic volatility model enhances its fit to the market implied volatility and improves its pricing performance in comparison to the addition of jump factors to the underlying stock price process.

References

[1] Alizadeh, S., Brandt, M.W. and Diebold, F.X. (2002). Range-Based Estimation of Stochastic Volatility Models. *Journal of Finance*, **57**, 1047 – 1091.

[2] Bakshi, G., Cao, C. and Chen, Z. (1997). Empirical Performance of Alternative Option Pricing Models. *Journal of Finance*, **52**, 2003 – 2049.

[3] Ball, C.A. and Roma, A. (1994). Stochastic Volatility Option Pricing. *Journal of Financial and Quantitative Analysis*, **29**, 589 – 607.

[4] Bates,D. (1996). Jump and Stochastic Volatility:Exchange Rate Processes Implicit in Deutche Mark in Options. *Review of Financial Studies* **9**, 69 – 107.

[5] Black, F. and Scholes, M. (1973). The Pricing of Options and Corporate Liabilities. *Journal of Political Economy*, **81**, 637 – 654.

[6] Christoffersen, P., Heston, S. and Jacobs, K. (2009). The Shape and Term Structure of the Index Option Smirk: Why Multifactor Stochastic Volatility Models Work So Well. *Management Science*, **55**, 1914 – 1932.

[7] Fouque, J.P. and Lorig, M.J. (2011). A Fast Mean-Reverting Correction to Heston's Stochastic Volatility Model. *SIAM Journal of Financial Mathematics*, **2**, 221 – 254.

[8] Fouque, J.P., Papanicolaou, G., Sircar, R. and Solna, K. (2003). Multiscale Stochastic Volatility Asymptotics. *Multiscale Modeling and Simulation*, **2**, 22 – 42.

[9] Fouque, J.P., Papanicolaou, G., Sircar, R. and Solna, K. (2011). *Multiscale Stochastic Volatility for Equity, Interest Rate and Credit Derivatives*. New York, USA: Cambridge University Press.

[10] Heston, S.L. (1993). A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options. *The Review of Financial Studies*, **6**, 327 – 343.

[11] Hull, J., & White, A. (1987). The Pricing of Options on Assets with Stochastic Volatilities. *Journal of Finance*, **42**, 281 – 300.

[12] Gatheral, J. (2006). *The Volatility Surface: A Practitioner's Guide*. Hoboken, NJ: Wiley.

[13] Kou, S.G. (2002). A Jump Diffusion Model for Option Pricing. *Management Science*, **48**, 1086 – 1101.

[14] Malhotra, G., Srivastava, R. and Taneja, H.C. (2018). Quadratic Approximation of the Slow Factor of Volatility in a Multifactor Stochastic Volatility Model. *Journal of Futures Markets*, **38**, 607 – 624.

[15] Merton,R.C. (1976). Option Pricing when Underlying Stock Returns are Discontinuous. *Journal of Financial Economics*. **3**, 125 – 144.

[16] Pan, J. (2002). The Jump-Risk Premia Implicit in Options: Evidence from an Integrated Time-Series Study. *Journal of Financial Economics*, **63**, 3 – 50.

[17] Scott, L.O. (1987). Option Pricing When the Variance Changes Randomly: Theory, Estimation, and an Application. *Journal of Financial and Quantitative Analysis*, **22**, 419 – 438.

[18] Scott, L.O. (1997). Pricing Stock Options in a Jump-Diffusion Model with Stochastic Volatility and Interest Rates: Applications of Fourier Inversion Methods. *Mathematical Finance*, **7**, 413 – 426.

[19] Shu, J. and Zhang, J.E. (2004). Pricing S&P 500 Index Options under Stochastic Volatility with the Indirect Inference Method. *Journal of Derivatives Accounting*, **1**, 171 – 186.

[20] Stein, E.M. and Stein, J.C. (1991). Stock Price Distributions with Stochastic Volatility: An Analytic Approach. *The Review of Financial Studies*, **4**, 727 – 752.

[21] Wiggins,J. (1987). Option Values under Stochastic Volatilities. *Journal of Financial Economics*, **19**, 351 – 372.

[22] Yan, G. and Hanson, F.B. (2006). Option Pricing for a Stocahstic Volatility Jump-Diffusion Model with Log-Uniform Jump Amplitudes. *Proceedings of the 2006 American control conference,USA*, 2989 – 2994.

Environmental Science and Pollution Research
https://doi.org/10.1007/s11356-019-07110-w

ENVIRONMENTAL AND ENERGY MANAGEMENT

Comparison of fuel characteristics of hydrotreated waste cooking oil with its biodiesel and fossil diesel

Ankit Sonthalia [1,2] · Naveen Kumar [2]

Received: 20 September 2019 / Accepted: 18 November 2019
© Springer-Verlag GmbH Germany, part of Springer Nature 2019

Abstract

Compression ignition engines powered by diesel are the work horses of developing countries like India. However, burning fossil fuel causes a lot of air pollution and the depletion of fuel at an alarming rate. Fuels produced from biomass or wastes can partially substitute fossil diesel to decrease its consumption. One such feedstock is waste cooking oil (WCO) which can be easily converted into fuel for diesel engines. The hydrotreating process stands out among the methods available for converting WCO into fuel, since its properties are almost similar to fossil diesel with little or no oxygen content. In this study, the physico-chemical properties of the hydrotreated waste cooking oil (HVO), biodiesel of waste cooking oil, diesel and blends of HVO and diesel are compared. The blends were prepared by mixing 10%, 20%, 30%, 40% and 50% of HVO on volume basis in diesel. The evaporation rate and ignition probability of the fuel samples were found by using a hot-plate test setup. HVO had higher ignition probability than all the test sample. As the percentage of HVO increased in the test samples, the ignition probability increased. The Sauter mean diameter (SMD) of the samples was also found using Malvern Spraytec. The SMD of HVO was larger than diesel but smaller than biodiesel. The study shows that blends of HVO up to 30% are feasible for present use in diesel engines, as the viscosity (2.54, 2.59 and 2.62cSt) and calorific value (42.41, 42.29, 42.08 MJ/kg) of the three blends (10%, 20% and 30%) is close to diesel (2.51cSt and 42.58 MJ/kg). Also, due to high cetane index, neat HVO or blends having higher HVO content (> 30%) cannot be used in the existing engines as the engine power output may be affected. Therefore, to use these fuels, the engine needs to be modified which is not feasible for existing engines. The FTIR and GC-MS analysis shows that the HVO has low oxygen content and high amount of paraffins, whereas biodiesel of waste cooking oil has high unsaturation and high oxygen content.

Keywords Waste cooking oil · Hydrotreated vegetable oil · Sauter mean diameter · Ignition probability · Evaporation rate

Introduction

Humans need energy to do any kind of work. The energy can be in the form of electricity or fuel for transport. With the increase in human population and change in lifestyle, the demand for energy has increased over the years. Fossil fuels meet this increase in energy demand; however, they are depleting at a fast rate and are also degrading the atmosphere by emitting greenhouse gas (Kumar and Sidharth 2018; Pali and Kumar 2016b). For the transportation and agricultural sector and power generation sector, diesel engines are extensively used (Ahlawat et al. 2013). However, the NOx and smoke emissions are high from the combustion of diesel in these engines which causes bronchitis and cancer (Machalaba et al. 2015; Pali and Kumar 2016a; Vermeulen et al. 2014). Therefore, reducing the emissions is necessary from the exhaust of a diesel-fuelled compression ignition engine.

The improvement in fuel efficiency and reduction in emissions simultaneously can be achieved either by modifying the engine or by the use of alternate fuels. Modifying the engine is not sustainable as it is costly and labour-intensive. Also modification of existing engines on-field is more difficult; therefore, only new engines can be modified. Another way is to use fuels derived from renewable sources which are environmentally benign. The vegetable oil can be transformed into biodiesel via

Responsible editor: Philippe Garrigues

✉ Ankit Sonthalia
ankit_sont@yahoo.co.in

[1] Department of Automobile Engineering, SRM Institute of Science and Technology, NCR Campus, Ghaziabad 201204, India

[2] Center for Advanced Studies and Research in Automotive Engineering, Delhi Technological University, Bawana Road, Rohini, Delhi 110042, India

Published online: 17 December 2019

transesterification reaction route which is a good alternative to diesel (Alwi et al. 2019; Damanik et al. 2018). The technology has matured over the years and is commercially used for biodiesel production. However, oxygen's presence in biodiesel reduces its heating value and decreases its oxidation stability (Li et al. 2018). Also, a lot of waste water is generated while washing the transesterified oil which is unsustainable from environment perspective (Sankaranarayanan et al. 2011). Researchers are upgrading the bio-oils to more stable hydrocarbons which are of high quality. The fuel is produced using specific catalysts via hydro-deoxygenation (HDO), hydro-decarbonylation (HDCN) or hydro-decarboxylation (HDCX) route (Mortensen et al. 2011; Saidi et al. 2014; Sonthalia and Kumar 2017). The produced fuel is linear containing less amount of sulphur, nitrogen and oxygen. The cetane number of fuel is high having good cold properties and high heating value. No separate infrastructure is required for storing, transporting and dispensing the fuel (Bhuiya et al. 2016).

A heterogeneous catalyst such as Co-Mo/Al_2O_3, Ni-Mo/Al_2O_3 and NiW/Al_2O_3 along with high temperature and pressure is required for the process. The oxygen content from the vegetable oil is removed and C_{15}-C_{18} range carbon atoms are produced. Co-processing the vegetable oil with heavy oil and gas oil results in the removal of nitrogen and sulphur as well (Demirbas and Dincer 2008; Kiss et al. 2006, 2008; Rajesh et al. 2015, 2016a, b; Zarchin et al. 2015). The effect of liquid hourly space velocity, hydrogen to oil ratio and pressure on the production of green diesel from waste cooking oil (WCO) using Ni-Mo catalyst was studied (Bezergianni et al. 2011). The fuel production was favourable at low pressures as high pressure promotes cracking. Reduction in the formation of green diesel is seen with the increase in hydrogen to oil ratio. Moreover, green diesel production is increased with increase in space velocity. The same research group (Bezergianni et al. 2010) also studied the effect of temperature on yield, selectivity and saturation of double bonds using a commercial hydrotreating catalyst. At low temperatures, the diesel like fuel production improved whereas at high temperatures, cracking occurred. The double bonds were also saturated at high temperatures. Three different types of catalysts namely a severe hydrocracking catalyst, a mild-hydrocracking catalyst and a hydrotreating catalyst were compared at various operating conditions by the same research group (Bezergianni et al. 2012). High conversion and diesel selectivity were observed with all the catalysts. Moreover, hydrotreating catalyst resulted in the highest removal of nitrogen and sulphur along with 99% saturation of fuel. Co-processing of waste cooking oil and heavy gas oil mixed in various ratios was studied (Bezergianni and Dimitriadis 2013). The effect of temperature on various parameters such as conversion, heteroatom removal, pour point, saturation of double bonds and hydrogen consumption was studied. Increase in temperature favours heteroatom (sulphur, nitrogen, oxygen etc.) removal, whereas the WCO content does not affect it. The content of WCO favours conversion whereas temperature does not significantly favour conversion. The pour point decreases with increase in WCO content and temperature.

Waste cooking oil was selected as the feedstock since the world produces billions of gallons of waste oil whose disposal and storage are costly (Hribernik and Kegl 2009). In the developing countries, the oil is thrown into the sewer system which pollutes the underground water, rivers and oceans. If the oil is converted into fuel, then two problems can be solved simultaneously. First: reduction in the cost of treating the wastes containing used cooking oil and reduction in the pollution of the environment. Second: the conversion of used oil into fuel can reduce the energy crisis (Arslan and Ulusoy 2018).

In the present work, waste cooking oil was hydrotreated to produce diesel like fuel. Also, biodiesel was produced from the waste cooking oil by the transesterification route. The physico-chemical properties of the two fuels were compared with the base diesel. The hydrotreated waste cooking oil was blended with diesel in various proportions and comparison was made with the neat fuels for evaluating their suitability as a compression ignition (CI) engine fuel. FTIR and GC-MS analysis of the neat fuels was compared. Ignition probability, evaporation time and microscopic spray characteristics of the test blends were also compared.

Production of waste cooking oil biodiesel

The waste cooking oil used in the present work was obtained from the cafeteria of Delhi Technological University. The oil had 0.95% free fatty acid content; therefore, the used cooking oil was converted into its biodiesel using a single-step transesterification process as previously done by Karmakar et al. (2012). Neat waste cooking oil biodiesel was designated as B100.

Production of the hydrotreated waste cooking oil

The hydroprocessing of waste cooking oil was carried out in an autoclave reactor. The reaction was carried out at 330 °C with an initial hydrogen pressure of 4 MPa. A 5% mix of ruthenium-, cerium- and manganese-based catalyst supported on alumina was used for the reaction. The weight of the catalyst was 2.5% of the waste cooking oil weight. After 200 min of reaction, the bio-crude yield was nearly 94%. The obtained bio-crude was then distilled into fractions according to ASTM D2892 and ASTM D5236 standards. The liquid fuel obtained in the boiling range of 180 to 370 °C was considered as green diesel or hydrotreated vegetable oil.

Blends preparation

The following blends were prepared with HVO and diesel in the present work: H10D90, H20D80, H30D70, H40D60 and H50D50. H10D90 contains 90% by volume diesel and 10% by volume HVO. Similarly, H20D80, H30D70, H40D60 and H50D50 contain 80, 70, 60 and 50% by volume diesel and 20, 30, 40 and 50% by volume HVO, respectively. Rigorous mixing was carried out to prepare the blends using splash blending process. The blends were kept in a cool and dry place for a period of 3 months and no phase separation was observed after 3 months. Table 1 tabulates the equipment used for the determination of the blends' properties.

Setup for hot plate ignition experiment

A stainless steel plate having 8 mm thickness and 152 mm diameter was selected. To heat the plate, a nichrome wire electrical heater was used. Using K-type thermocouple, the plate's temperature was measured and the electrical supply to the heater was varied using a variac to change the temperature of the plate. As shown in Fig. 1, the plate was placed directly above the heater and the empty space was filled with glass wool to minimize the heat loss to the surroundings. The plate surface was polished finely and the centre of the plate had a small concave depression so that the droplet stays in the middle of the plate. The fuel droplet was dropped on plate's centre using a firmly mounted pipette. The pipette was placed 20 mm above the steel plate so that the flame of the droplet on the hot-plate does not ignite the fuel in the pipette and the fuel droplet does not break while falling. At a particular temperature, a single droplet of fuel was dropped. To ascertain that the droplet size was the same for all the experiments, the average size of the droplets was determined by finding the number of droplets generated for a fixed volume. The average droplet size was 2.6 mm and the size was same for all the test samples.

After every experiment, glazing of the plate was carried out using sand paper. For each test fuel, ten tests were carried out and for each surface temperature, three readings were taken. For finding the evaporation time, the temperature was varied between 150 and 500 °C. For finding the ignition probability, two possible events, namely no ignition of droplet and ignition of droplet, were recorded. At a particular temperature, a droplet was dropped ten times for finding out the ignition probability. The temperature was varied from 280 to 640 °C in steps of 20 °C. For the experiment, different blends of HVO and diesel and neat diesel and biodiesel samples were used. A digital camera with 90 frames per second speed was used for recording the droplet evaporation and burning process. The time interval of the process was measured with 0.01-s accuracy. Time was measured from the moment the drop landed on the plate till it evaporated. A room under atmospheric conditions was used for carrying out the experiments. The uncertainty of the evaporation time experiment was calculated according to the method given by Kline and McClintock (Kline and McClintock 1953) and was found to be ± 3%.

Setup for spray characteristics

A Malvern make phase Doppler particle analyzer (PDPA) was used in the present work. Table 2 shows the specification of the Malvern make (Spraytec) instrument. The PDPA is based upon the Laser Doppler effect wherein the spherical particles are passed through a laser beam and the scattering pattern is predicted. The size and particle concentration is determined through the scattering pattern. The fuel was sprayed through a three-hole injector in which two holes were closed and the fuel was sprayed through a single hole at 200 bar injection pressure. The test fuel was supplied to the injector through a common rail and a high pressure pump. Figure 2 shows the experimental setup. For each blend, five readings were taken and averaged. The distance between the injector and the laser beam was fixed at 60 mm. The data was collected for 200 ms although a single injection lasted about 12 ms. The data reported in this work is displayed at 6 ms, so that a clear comparison can be made of the droplet distribution in the middle of the injection. The tests were carried out in ambient conditions in a low-light room so that the effect of sunlight was minimum. The disintegration of the fuel jet is dependent upon the properties of the fuel, the injector's geometry and the density of the atmosphere where this disintegration takes place, called primary atomization. At the time of fuel injection

Table 1 Equipment and test standard for determining properties of blends

Property	Equipment	Test standard
Density	Anton Parr DMA 4500	ASTM D4052
Kinematic viscosity	Capillary tube	ASTM D445
Calorific value	Parr 6100 Oxygen Bomb Calorimeter	ASTM D240
Flash point	Pensky-Martens (AD0093–600)	ASTM D93
Cetane index	–	ASTM D4737
CFPP	NEWLAB 200/1	ASTM D6731
GC-MS	QP-2010 plus	ASTM D3973
HFRR	HFRR test rig	ASTM D6079

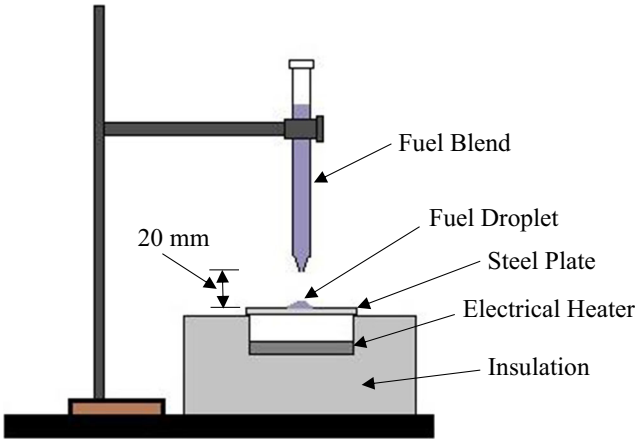


Fig. 1 Schematic of hot plate test setup

in a diesel engine, the cylinder pressure is about 30–40 bar which will cause secondary atomization of the fuel followed by primary atomization (Daho et al. 2013). The present study aims to investigate the effect of only fuel properties on the fuel spray; therefore, primary atomization was studied by carrying out experiments in ambient conditions. The uncertainty of the Sauter mean diameter was also calculated according to the method given by Kline and McClintock (Kline and McClintock 1953) and was found to be ± 1.3%.

Results and discussion

FTIR and GC-MS analysis of HVO and biodiesel

FTIR is used for analysing the atoms or groups in a molecule and the structure of molecules based upon the energies associated with the molecular vibrations. In the present study, a Thermo-Fisher Scientific make, Nicolet 380 model, was used to carry out the FTIR analysis. Initially, a spectrum of the optical path was obtained without the sample to understand the background and then the sample was placed in the infrared beam path to find its spectrum. The spectrum with the sample and the background were co-added and then Fourier-transformed by the software thereby producing the desired FTIR spectrum.

Generally, the C–O stretching is observed in the range of 1100–1300 cm^{-1} which is present in the methyl esters, ethers, carboxylic acids and alcohols (Parada Hernandez et al. 2017). The methyl ($–CH_3$) stretching vibration is observed in the range of 1375–1450 cm^{-1}. The bending vibration of methylene group is seen at 1465 cm^{-1} (scissoring vibration) and 721 cm^{-1} (rocking vibration) (Mahesh et al. 2015). The stretching of –C=O is obtained in the range of 1730–1750 cm^{-1}. The aryl and α, β unsaturated structure may cause this band to shift slightly towards 1715–1730 cm^{-1} (Matwijczuk et al. 2017; Omidvarborna et al. 2016). The alkanes stretch is observed from 2850 to 3000 cm^{-1}. The Fourier-transform infrared spectrum of hydrotreated used cooking oil and the biodiesel of used cooking oil is compared in Fig. 3. It is seen that the biodiesel has a peak at 1742 wave number which represents the carboxyl group. After hydrogenation of oil, the peak lost its intensity. During hydrogenation, the hydrogen reacts with the oil in the presence of the catalyst and the oxygen present in the triglyceride is removed as water resulting in the formation of hydrocarbons. Similar results were also observed by Satyarthi and Srinivas (Satyarthi and Srinivas 2011). The C–H stretching peaks were observed at 2854 cm^{-1} and 2924 cm^{-1} wave number in biodiesel. It is seen that the intensity of the peaks have increased with HVO, this is mainly due to the conversion of the esters into hydrocarbons.

Figures 4 and 5 shows the GC-MS profile of WCO biodiesel and HVO, respectively. The GC-MS of the samples was carried out using an offline Shimadzu QP-2010 instrument equipped with Agilent DB-2887 column (10 m × 0.53 mm × 3.0 μm) and a flame ionization detector. The liquid samples of 0.1 μl were injected into the

Table 2 Specifications of Malvern Spraytec

Measurement principle	Laser diffraction
Size range	0.1–2000 μm
Optical models	Mie theory and Fraunhofer approximation including patented multiple scattering correction
Lens range	750 mm lens: 2.0–2000 μm (D_{v50}, 5–1600 μm)
Working range	100 mm at 0.5 μm, > 1 m at 10 μm
Concentration range	Minimum acceptable transmission, 5%
Detection system	35 element log-spaced silicon diode detector array
Light source	632.8 nm, 5-mW helium–neon laser
Maximum acquisition rate	Continuous mode, 1 Hz
	Rapid mode, 2.5 kHz as standard, 10 kHz with additional software feature key
Accuracy	Better than ± 1% on the D_{v50}
Precisions/repeatability	Better than ± 1% COV on the D_{v50}

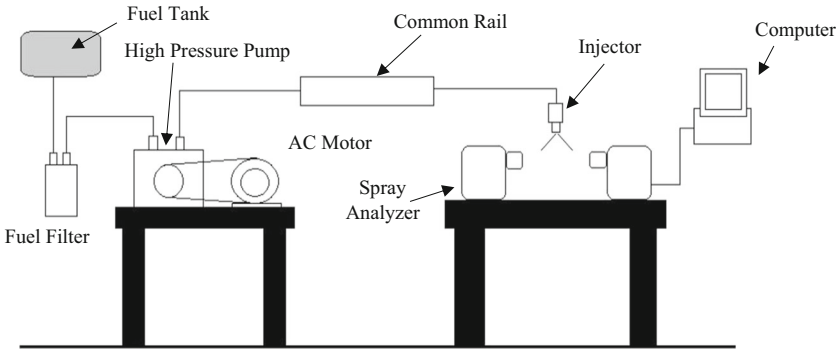


Fig. 2 Schematic of the spray analysis setup

column with the detector and injector temperature maintained at 350 °C. The GC oven temperature was controlled by a programmable temperature controller which had a ramp rate of 15 °C/min. Nitrogen was used as the inert gas for carrying the liquid samples through the column. Figure 4 shows that the biodiesel predominantly contains palmitic acid, oleic acid and methyl linoleate. The carbon atoms in the biodiesel ranged from C-15 to C-21. The cetane index of biodiesel is 43.78 and it has a kinematic viscosity of 5.82cSt. The GC-MS profile of HVO (Fig. 5) shows that the carbon atoms range from C-11 to C-18 and it predominantly contains heptadecane. Its cetane index is 70.2. In addition, methyl esters are also observed that shows the hydrotreating process was incomplete. The FTIR of the test fuel (Fig. 3) confirms the same. It can be inferred from the results that HVO contains normal alkanes that tends to improve the ignition quality which increases with increase in chain length (Ogunkoya et al. 2015). Also, as the chain length of the hydrocarbons increases, the energy density and the volatility increase (Bacha et al. 2007). Therefore, the combustion with HVO will be much better than diesel and biodiesel.

Evaporation time and ignition probability

Figure 6 portrays the evaporation of the test fuels as a function of the plate temperature. The trends of the test fuels are similar and they can be divided into four broad regions: evaporation, nucleate boiling, transition boiling and film boiling (Bennett 2001). In the evaporation region, the droplet is in direct contact with the hot plate. As the surface temperature increases, the evaporation time decreases since the heat transfer rate through conduction increases. In the nucleate boiling region, bubbles were formed on the surface of the hot plate. The fluid mixing near the surface increases due to the bubbles formed and as the temperature increases the heat transfer rate increases thereby decreasing the evaporation time. The bubble formation started for D100, H100 and B100 at 250, 195 and 280 °C, respectively. The reason for the trend is the boiling time of liquids which depends upon their boiling temperature and heat capacity. Hence, liquid with the shortest boiling temperature will evaporate first and vice versa (Cengel and Turner 2017). It is also seen that as the percentage of HVO increases in diesel, the evaporation time decreases, since HVO evaporates first and then the heavier components are evaporated.

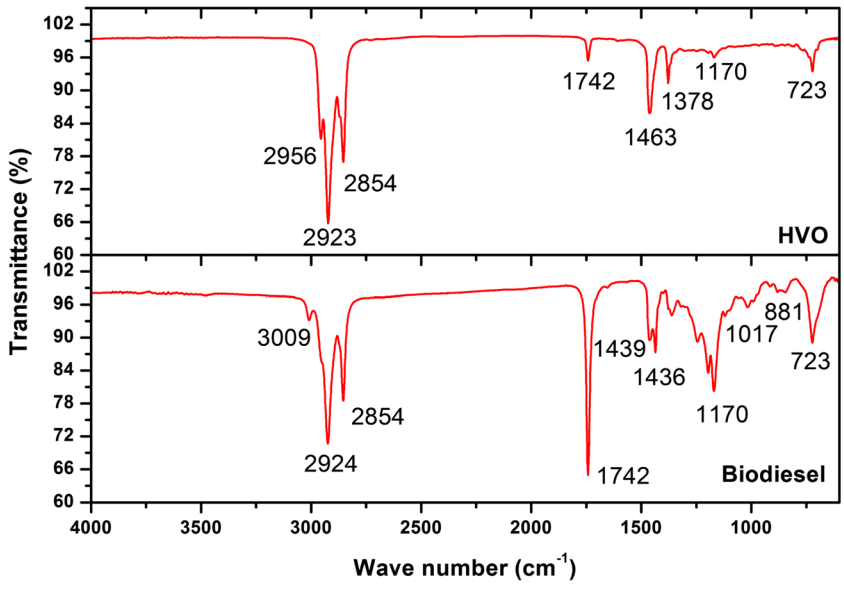


Fig. 3 Infrared spectra of HVO and biodiesel

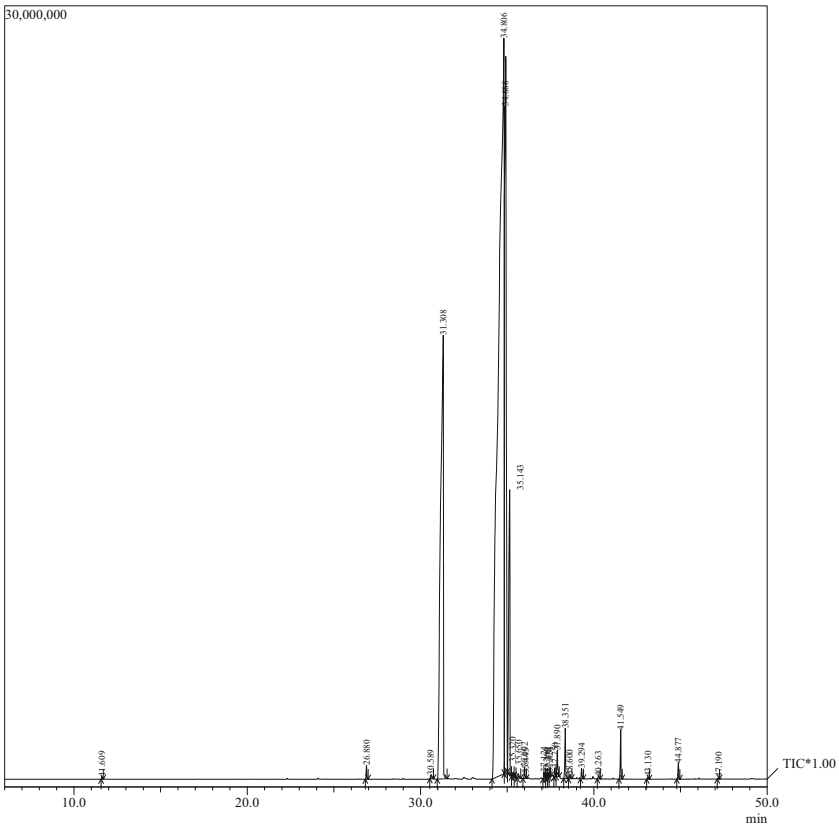


Fig. 4 GC-MS profile of waste cooking oil biodiesel

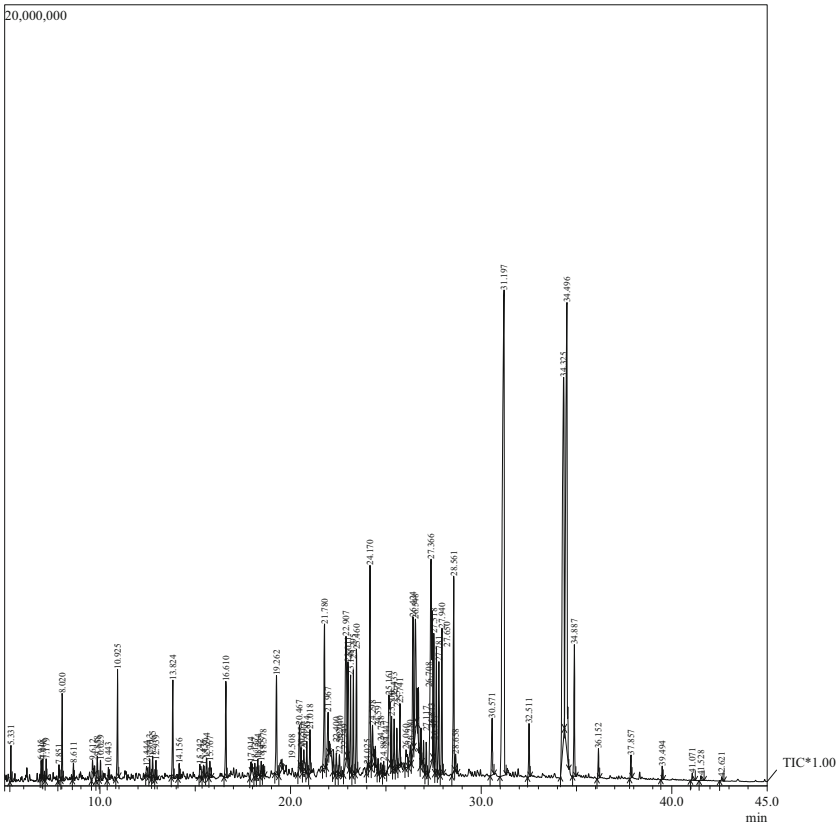


Fig. 5 GC-MS profile of HVO

The evaporation time is lowest at a certain surface temperature called the critical temperature. The droplets at this temperature disintegrate after significant boiling has occurred, resulting in almost instantaneous evaporation of the test fuels. The critical temperature for D100, H100 and B100 is 360, 270 and 378 °C, respectively. Abu-Zaid (2004) also got similar results. The critical temperature of HVO is observed to be lowest. Also, as the HVO percentage in the blends with diesel increases, the critical temperature reduces. With further increase in temperature, a transition boiling region is observed. In this region, a vapour film starts forming around the droplet due to excessive formation of bubbles and the direct contact between the droplet and the hot surface is broken. Since the vapour film has a lower thermal conductivity than that of liquid, the heat transfer between the two reduces thereby increasing the evaporation time. In the film boiling region, Leidenfrost phenomena is observed, that is the droplet surrounded by the vapour layer starts levitating over the surface of the hot plate. As the surface temperature increases, the heat transfer increases since the transfer of heat takes place through thermal radiation and conduction resulting in decrease in evaporation time. In this region, the latent heat of vaporization plays a major role (Cengel and Turner 2017). Biodiesel has the highest evaporation time in this region, and H100 has the lowest evaporation time. The evaporation time with diesel was observed to be lower than B100. The blends of HVO and diesel had evaporation time in between the two test fuels.

The ignition probability of the test fuels with respect to the surface temperature is shown in Fig. 7. The ignition probability was calculated using logistic regression and the fit quality was determined using chi-squared goodness fit test (Somandepalli et al. 2008). It is seen that the ignition probability of H100 is the highest at all the temperatures and 100% probability is observed at 560 °C. The lowest ignition probability is observed with B100 followed by diesel. The blends of diesel and HVO had probabilities in between the two neat fuels. Previous studies (Arifin and Arai 2010; Arifin et al. 2008) show that the ignition probability is more in the film boiling region. In the film boiling region, a vapour film covers the droplet and heat transfer takes place through conduction and radiation. As the surface temperature increases, the radiation through the vapour increases, thereby increasing the ignition probability. Therefore, in this work, the ignition probability tests were conducted in between 280 and 640 °C as the film boiling occurs in this region (Fig. 6). It is seen that the density and viscosity of the fuel play major roles in the ignition probability (Shams and Moghiman 2018). Since biodiesel has the highest density and viscosity, its ignition probability is the lowest. H100 has low density and high viscosity compared with diesel; still the ignition probability is high with H100 due to its lower boiling temperature.

Microscopic spray characteristics of the test fuels

The fuel spray in terms of particle size plays an important role in the quality of the combustion. It is a known fact that the smaller the particle size, the greater will be its vaporization and the early burning tendency of the fuel will be higher. However, finer particles have low momentum and their penetration distance is less resulting in incomplete utilization of air. If the particle size is large, then its vaporization will be slow and the delay in its ignition may be high resulting in more heat released in the later stages of combustion (diffusion combustion phase) that tends to increase the exhaust temperature and the utilization of the fuel energy for useful work is less. Therefore, the particle size should be optimum for smooth and efficient engine running (Heywood 1988).

The particle size of a fuel spray cannot be directly evaluated; therefore, the average diameter of the droplet is statistically calculated and it is called Sauter mean diameter (SMD). SMD is the

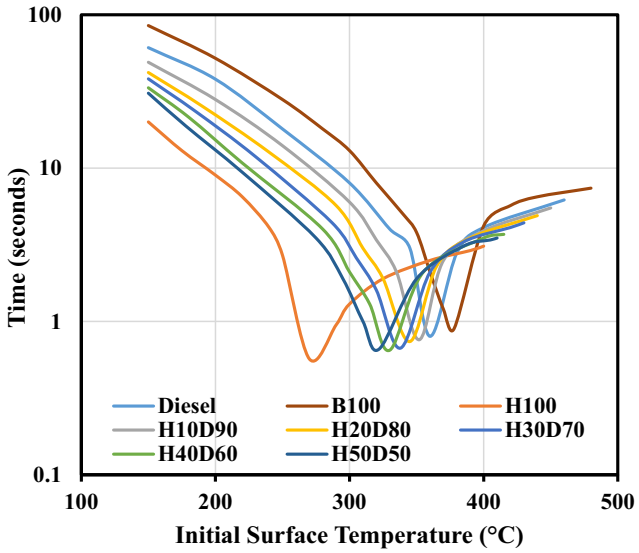


Fig. 6 Variation of evaporation time with surface temperature of plate

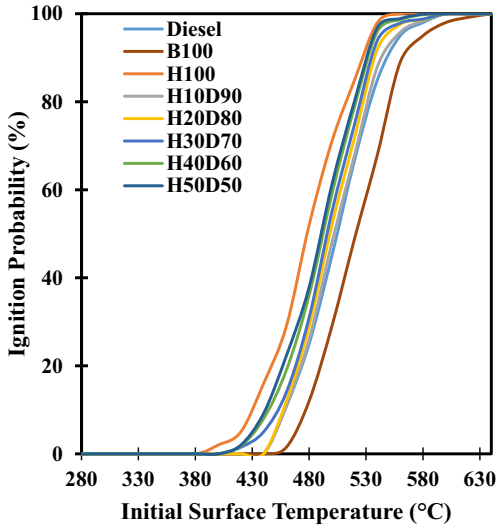


Fig. 7 Variation of ignition probability with surface temperature of plate

mean diameter of the spherical particles having the same volume to surface ratio. SMD is dependent upon the fuel properties such as viscosity, density and surface tension. It is also dependent upon the fuel injection pressure; however, in this study, the fuel injection pressure was fixed at 200 bar, since the engine for which this study is being conducted operates at this pressure. Figure 8 shows the SMD of neat diesel, HVO and biodiesel and blends of HVO and diesel. It is seen that neat biodiesel has the highest SMD followed by HVO and the lowest SMD is observed with neat diesel. Table 3 shows that the biodiesel has the highest viscosity and density resulting in larger particle size during the spray. The instabilities in the fuel jet required to break up into small particles are adversely affected when its viscosity is high. Also, due to high density, the atomization is affected and the droplet formation tendency is opposed due to the fuel's high surface tension (Ejim et al. 2007). Compared with biodiesel, HVO has low density and viscosity whereas its density is lower than diesel but its viscosity is higher. Due to the high viscosity of HVO, the breakup of the fuel jet is improper resulting in slightly higher SMD than neat diesel. Figure 8 also shows that 10% (by volume) of HVO in the blend has SMD close to diesel as the fuel has properties similar to diesel. With the increase in HVO percentage in the blend, the SMD increases.

The specific surface area (SSA) of the test fuels was also calculated by the Malvern Spraytec software (Table 2). Figure 9 shows the SSA for the test fuels. Specific surface area is the ratio of the surface area of the particle to the volume of the particle. The higher the specific surface area, the more will be the particles surrounded by the oxygen molecules, the lesser will be the delay in ignition resulting in good combustion (Kumar and Tomar 2019). Among the neat fuels, diesel had the highest surface area followed by HVO and waste cooking oil biodiesel. The particles that are small in size have a high surface area. As seen in Fig. 8, it is clear that diesel has the smallest particle size; therefore, its surface area is the largest. Also, 10% blend of HVO in diesel has surface area close to diesel and the SSA keeps on decreasing as the percentage of HVO in the blend increases.

Discussion on properties of neat diesel, biodiesel, HVO and blends of HVO and biodiesel

In this section, the feasibility of using the fuels in a compression ignition engine is discussed on the basis of the physico-chemical properties of the fuels. Table 3 shows the properties of the fuels. It is seen that biodiesel has the highest density followed by diesel and HVO. The kinematic viscosity of biodiesel is the highest, whereas HVO has viscosity in between diesel and biodiesel. It is a well-known fact that after the injection of the fuel in the combustion chamber, its atomization and mixing with air are affected by its viscosity and density. The higher the viscosity and density, the lower is the tendency of the fuel to mix with air resulting in improper combustion and higher emissions (Boopathi et al. 2017; Geo et al. 2017; Varuvel et al. 2018). The blends of HVO and diesel tend to have reduced the density and increased the viscosity with the increase of HVO in the blend. The presence of oxygen increases the viscosity of fuel and the presence of higher number of hydrocarbons decreases the density (Manchanda et al. 2018). As seen in the FTIR analysis of the HVO and biodiesel, the oxygen content of HVO is lower than that of biodiesel whereas diesel does not contain oxygen; this explains the reason for the trend. Also, due to lower oxygen content, the calorific value of HVO is higher than biodiesel but

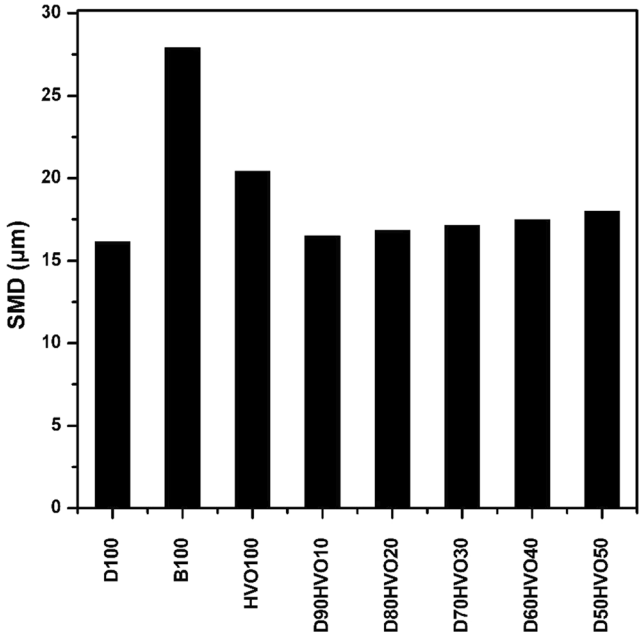


Fig. 8 Sauter mean diameter of the test fuels

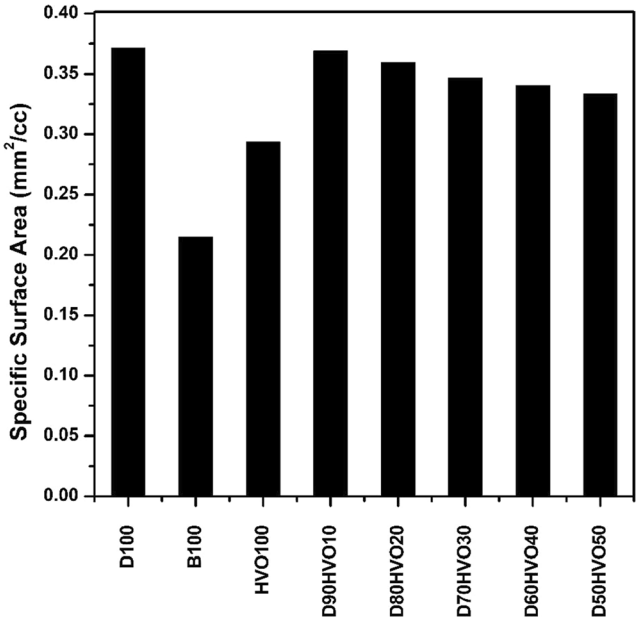


Fig. 9 Specific surface area of the test fuels

Table 3 Properties of the neat diesel, HVO and their blends

Property		D100	B100	H100	H10D90	H20D80	H30D70	H40D60	H50D50
Density at 15 °C (kg/m^3)		835.07	863.71	817.7	832.4	829.68	829.1	828.5	827.19
Kinematic viscosity at 40 °C (cSt)		2.51	5.82	2.9	2.54	2.59	2.62	2.66	2.69
Calorific value (MJ/kg)		42.57	41.27	41.79	42.41	42.29	42.08	41.92	41.86
Flash point (°C)		58	165	68	59.5	60.2	61.1	62.3	62.8
Cetane index		45.73	46.93	65.2	47.95	49.93	51.0	51.2	53.9
Cloud point (°C)		− 0.8	4	− 6	− 1.1	− 1.6	− 2.3	− 2.8	− 3.2
CFPP (°C)		− 11	− 6	− 13	− 11.2	− 11.6	− 11.8	− 12	− 12.2
Lubricity, HFRR (microns)		370.2	175	194	363.4	349.6	324.1	303.6	280.8
Distillation Data (°C)	10%	152	220	196	158	163	170	172	180
	50%	248	310	293	254	258	261	263	269
	90%	338	340	322	334	331	329	327	324

its calorific value is lower than diesel. With the increase in HVO content in the blends, the calorific value was found to decrease. The cetane index is a measure of the ignition quality of fuel. The higher the cetane index, the lesser is the delay in ignition and vice versa (Heywood 1988). The cetane index was calculated by using ASTM D4737 standard for which the distillation tests of the samples was carried out. It is seen that the cetane index of HVO is higher than diesel and biodiesel. Figure 7 confirms that the ignition of neat HVO will be earlier than diesel and biodiesel, since the ignition probability of HVO is higher than diesel and biodiesel at a particular temperature and it reaches 100% probability at a lower temperature than the two fuels. The cold flow properties of biodiesel i.e. cloud point, CFPP are poor in comparison with diesel and HVO. Since HVO has higher isomeric content, its cloud point and CFPP are lower than diesel (Bezergianni et al. 2014). As the percentage of HVO in the blend increases, its cloud point and CFPP start decreasing. This shows that the fuel can be used in winters without any problems of filter clogging. The lubricity of the fuel for use in diesel engines is a very important property as the fuel pump is lubricated by the fuel only. Table 3 shows that the lubricity of biodiesel is the highest as the scar diameter during an HFRR test was smallest, whereas HVO has lubricity in between the biodiesel and diesel. The lubricity of a fuel is dependent upon the presence of esters in a fuel. Since biodiesel is a fatty acid ester, its lubricity is high. Similar results were found by Singh et al. Singh et al. (2018).

The blends and neat HVO and diesel were also tested in a single-cylinder direct-injection diesel engine. The tests were carried out at 25%, 50%, 75% and full load. Figure 10 shows the NO-smoke trade-off for all the samples. In general, there is a trade-off between the NO and smoke emissions in a conventional diesel engine. The trade-off exists since the presence of higher amount of oxygen tends to increase the NO emission whereas the deficiency of oxygen tends to increase the smoke emission (Varuvel et al. 2018). It is seen that the smoke opacity and NO emissions were the highest for engine operation with diesel. As the percentage of HVO increases in the blend, both the smoke and NO emission decrease and lowest emissions were observed with H100. As HVO has a higher cetane number, the ignition of the injected fuel started earlier than diesel and due to low calorific

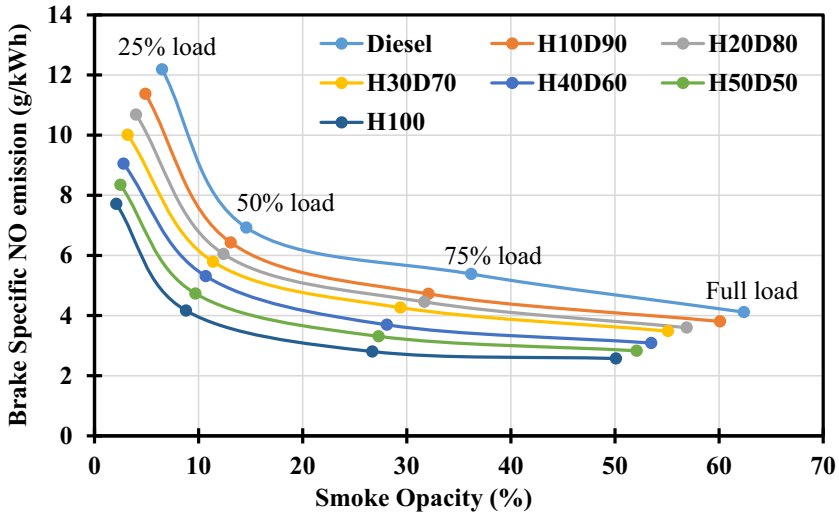


Fig. 10 NO-smoke trade-off for the test samples at different loads

value of the fuel, the amount of heat released was lower than diesel thereby reducing the combustion temperature. The reduced combustion temperature helped in lowering the NO emissions. The aromatic content of the fuel plays a major role in the soot formation, since the HVO has almost negligible aromatic content the smoke emissions reduced. Although not shown in this study, the HC and CO emissions were also measured for the test samples and it was seen that both the HC and CO emissions were lower than diesel till 30% blend of HVO in neat diesel. Further increase in the percentage of HVO in the blend resulted in an increase in both the emissions and the highest emissions were observed with neat HVO engine operation.

Conclusions

Based on the results, it is clear that neat HVO has properties better than biodiesel but similar to diesel. Its evaporation rate and ignition probability are higher than those of both diesel and biodiesel. Its SMD is slightly higher than that of diesel on account of its higher viscosity. It is seen that its cold flow properties are better than those of diesel and its lubricity is also better than that of diesel. However, at present, the use of neat HVO as a fuel in the diesel engine is not recommended. Firstly, the viscosity of HVO is high and its calorific value is lower than that of diesel. Secondly, the cetane index of HVO is high; therefore, some engine parameters need to be adjusted for utilizing the full potential of the fuel. Thirdly, the engine test results show that both smoke and NO emissions can be reduced with the use of HVO in a diesel engine. However, the lowest HC and CO emissions were observed with H30D100. With the increase in percentage of HVO in the blend, the HC and CO emissions increased. Lastly, it is impossible to completely replace diesel since the current world consumption of the fuel is so high that all the used cooking oil in the world will not be enough for producing the fuel. Therefore, the blends having HVO up to 30% in a litre of sample are advisable to use in a diesel engine.

References

Abu-Zaid M (2004) An experimental study of the evaporation characteristics of emulsified liquid droplets. Heat Mass Transf 40(9):737–741. https://doi.org/10.1007/s00231-003-0473-5

Ahlawat V, Gupta M, Anand S, Bansal V, Jain V, Kumar N (2013) Evaluation of performance and emission characteristics of an unmodified naturally aspirated compression ignition engine on blends of diethyl ether and diesel. In 8th SAEINDIA International Mobility Conference & Exposition and Commercial Vehicle Engineering Congress 2013 (SIMCOMVEC). SAE International. https://doi.org/10.4271/2013-01-2888

Alwi A, Zulkifli NW, Sukiman NL, Yusoff A, Zakaria Z, Jamshaid M, Hasnul MH, Amzar MS (2019) Evaluation of engine performance and exhaust emission characteristics in a diesel engine using isobutanol—Calophyllum inophyllumbiodiesel—diesel ternary blends. Environ Sci Pollut Res 26(12):11815–11826. https://doi.org/10.1007/s11356-019-04603-6

Arifin YM, Arai M (2010) The effect of hot surface temperature on diesel fuel deposit formation. Fuel 89(5):934–942. https://doi.org/10.1016/j.fuel.2009.07.014

Arifin YM, Furuhata T, Saito M, Arai M (2008) Diesel and bio-diesel fuel deposits on a hot surface. Fuel 87(8):1601–1609. https://doi.org/10.1016/j.fuel.2007.07.030

Arslan R, Ulusoy Y (2018) Utilization of waste cooking oil as an alternative fuel for Turkey. Environ Sci Pollut Res 25(25):24520–24525. https://doi.org/10.1007/s11356-017-8899-3

Bacha J, Freel J, et al. (2007)Diesel fuels technical review. San Ramon, CA. Retrieved from https://www.chevron.com/-/media/chevron/operations/documents/diesel-fuel-tech-review.pdf

Bennett M (2001) Ignition of combustible fluids by heated surfaces. Process Saf Prog 10(1):29–36

Bezergianni S, Dimitriadis A (2013) Temperature effect on co-hydroprocessing of heavy gas oil-waste cooking oil mixtures for hybrid diesel production. Fuel 103:579–584. https://doi.org/10.1016/j.fuel.2012.08.006

Bezergianni S, Dimitriadis A, Kalogianni A, Pilavachi PA (2010) Hydrotreating of waste cooking oil for biodiesel production. Part I: effect of temperature on product yields and heteroatom removal. Bioresour Technol 101(17):6651–6656. https://doi.org/10.1016/j.biortech.2010.03.081

Bezergianni S, Dimitriadis A, Kalogianni A, Knudsen KG (2011) Toward hydrotreating of waste cooking oil for biodiesel production, effect of pressure, H2/oil ratio, and liquid hourly space velocity. Ind Eng Chem Res 50(7):3874–3879. https://doi.org/10.1021/ie200251a

Bezergianni S, Kalogianni A, Dimitriadis A (2012) Catalyst evaluation for waste cooking oil hydroprocessing. Fuel 93:638–641. https://doi.org/10.1016/j.fuel.2011.08.053

Bezergianni S, Dimitriadis A, Chrysikou LP (2014) Quality and sustainability comparison of one- vs. Two-step catalytic hydroprocessing of waste cooking oil. Fuel 118:300–307. https://doi.org/10.1016/j.fuel.2013.10.078

Bhuiya MMK, Rasul MG, Khan MMK, Ashwath N, Azad AK, Hazrat MA (2016) Prospects of 2nd generation biodiesel as a sustainable fuel – part 2: properties, performance and emission characteristics. Renew Sust Energ Rev 55:1129–1146. https://doi.org/10.1016/J.RSER.2015.09.086

Boopathi D, Sonthalia A, Devanand S (2017) Experimental investigations on the effect of hydrogen induction on performance and emission behaviour of a single cylinder diesel engine fuelled with palm oil methyl ester and its blend with diesel. J Eng Sci Technol 12(7):1972–1984

Cengel Y, Turner R (2017) Fundamentals of thermal fluid sciences (fourth). Mc-Graw Hill Education, New York

Daho T, Vaitilingom G, Ouiminga SK, Piriou B, Zongo AS, Ouoba S, Koulidiati J (2013) Influence of engine load and fuel droplet size on performance of a CI engine fueled with cottonseed oil and its blends with diesel fuel. Appl Energy 111:1046–1053. https://doi.org/10.1016/j.apenergy.2013.05.059

Damanik N, Ong HC, Tong CW, Mahlia TMI, Silitonga AS (2018) A review on the engine performance and exhaust emission characteristics of diesel engines fueled with biodiesel blends. Environ Sci Pollut Res 25(16):15307–15325. https://doi.org/10.1007/s11356-018-2098-8

Demirbas A, Dincer K (2008) Sustainable green diesel: a futuristic view. Energy Sources Part A 30(13):1233–1241. https://doi.org/10.1080/15567030601082829

Ejim CE, Fleck BA, Amirfazli A (2007) Analytical study for atomization of biodiesels and their blends in a typical injector: surface tension and viscosity effects. Fuel 86(10–11):1534–1544. https://doi.org/10.1016/J.FUEL.2006.11.006

Geo VE, Sonthalia A, Nagarajan G, Nagalingam B (2017) Studies on performance, combustion and emission of a single cylinder diesel engine fuelled with rubber seed oil and its biodiesel along with ethanol as injected fuel. Fuel 209:733–741. https://doi.org/10.1016/j.fuel.2017.08.036

Heywood JB (1988) Internal combustion engine fundamentals. Mc-Graw Hill

Hribernik A, Kegl B (2009) Performance and exhaust emissions of an indirect-injection (IDI) diesel engine when using waste cooking oil as fuel. Energy Fuel 23:1754–1758

Karmakar A, Karmakar S, Mukherjee S (2012) Biodiesel production from neem towards feedstock diversification: Indian perspective. Renew Sust Energ Rev 16(1):1050–1060. https://doi.org/10.1016/J.RSER.2011.10.001

Kiss AA, Dimian AC, Rothenberg G (2006) Solid acid catalysts for biodiesel production –-towards sustainable energy. Advanced Synthesis & Catalysis 348(1–2):75–81. https://doi.org/10.1002/adsc.200505160

Kiss AA, Dimian AC, Rothenberg G (2008) Biodiesel by catalytic reactive distillation powered by metal oxides. Energy Fuel 22(1):598–604. https://doi.org/10.1021/ef700265y

Kline SJ, McClintock FA (1953) Describing uncertainties in single sample experiments. Mech Eng:3–8. https://doi.org/10.1111/jcmm.13453

Kumar N, Sidharth (2018) Some studies on use of ternary blends of diesel, biodiesel and n-octanol. Energy Sources Part A 40(14):1721–1728. https://doi.org/10.1080/15567036.2018.1486902

Kumar N, Tomar M (2019) Influence of nanoadditives on ignition characteristics of Kusum (Schleichera oleosa) biodiesel. Int J Energy Res 43(8):3223–3236. https://doi.org/10.1002/er.4446

Li X, Luo X, Jin Y, Li J, Zhang H, Zhang A, Xie J (2018) Heterogeneous sulfur-free hydrodeoxygenation catalysts for selectively upgrading the renewable bio-oils to second generation biofuels. Renew Sust Energ Rev 82(July):3762–3797. https://doi.org/10.1016/j.rser.2017.10.091

Machalaba C, Romanelli C, Stoett P, Baum SE, Bouley TA, Daszak P, Karesh WB (2015) Climate change and health: transcending silos to find solutions. Ann Glob Health 81(3):445–458. https://doi.org/10.1016/J.AOGH.2015.08.002

Mahesh SE, Ramanathan A, Begum KMMS, Narayanan A (2015) Biodiesel production from waste cooking oil using KBr impregnated CaO as catalyst. Energy Convers Manag 91:442–450. https://doi.org/10.1016/J.ENCONMAN.2014.12.031

Manchanda T, Tyagi R, Sharma DK (2018) Comparison of fuel characteristics of green (renewable) diesel with biodiesel obtainable from algal oil and vegetable oil. Energy Sources Part A 40(1):54–59. https://doi.org/10.1080/15567036.2017.1405109

Matwijczuk A, Zając G, Kowalski R, Kachel-Jakubowska M, Gagoś M (2017) Spectroscopic studies of the quality of fatty acid methyl esters derived from waste cooking oil. Pol J Environ Stud 26(6):2643–2650. https://doi.org/10.15244/pjoes/70431

Mortensen PM, Grunwaldt J-D, Jensen PA, Knudsen KG, Jensen AD (2011) A review of catalytic upgrading of bio-oil to engine fuels. Appl Catal A Gen 407(1–2):1–19. https://doi.org/10.1016/J.APCATA.2011.08.046

Ogunkoya D, Roberts WL, Fang T, Thapaliya N (2015) Investigation of the effects of renewable diesel fuels on engine performance, combustion, and emissions. Fuel 140:541–554. https://doi.org/10.1016/j.fuel.2014.09.061

Omidvarborna H, Kumar A, Kim D-S (2016) Variation of diesel soot characteristics by different types and blends of biodiesel in a laboratory combustion chamber. Sci Total Environ 544:450–459. https://doi.org/10.1016/J.SCITOTENV.2015.11.076

Pali HS, Kumar N (2016a) Combustion, performance and emissions of Shorea robusta methyl ester blends in a diesel engine. Biofuels 7(5):447–456. https://doi.org/10.1080/17597269.2016.1153363

Pali HS, Kumar N (2016b) Comparative assessment of Sal and kusum biodiesel properties. Energy Sources Part A 38(22):3391–3396. https://doi.org/10.1080/15567036.2015.1136974

Parada Hernandez NL, Bonon AJ, Bahú JO, Barbosa MIR, Wolf Maciel MR, Filho RM (2017) Epoxy monomers obtained from castor oil using a toxicity-free catalytic system. J Mol Catal A Chem 426:550–556. https://doi.org/10.1016/J.MOLCATA.2016.08.005

Rajesh M, Sau M, Malhotra RK, Sharma DK (2015) Hydrotreating of gas oil, Jatropha oil, and their blends using a carbon supported cobalt-molybdenum catalyst. Pet Sci Technol 33(19):1653–1659. https://doi.org/10.1080/10916466.2015.1036291

Rajesh M, Sau M, Malhotra RK, Sharma DK (2016a) Synthesis and characterization of Ni-Mo catalyst using Jatropha curcas leaves as carbon support and its catalytic activity for hydrotreating of gas oil, Jatropha oil, and their blends. Pet Sci Technol 34(3):240–246. https://doi.org/10.1080/10916466.2015.1124890

Rajesh M, Sau M, Malhotra RK, Sharma DK (2016b) Synthesis and characterization of Ni-Mo catalyst using pea pod (Pisum sativum L) as carbon support and its hydrotreating potential for gas oil, Jatropha oil, and their blends. Pet Sci Technol 34(4):394–400. https://doi.org/10.1080/10916466.2015.1120748

Saidi M, Samimi F, Karimipourfard D, Nimmanwudipong T, Gates BC, Rahimpour MR (2014) Upgrading of lignin-derived bio-oils by catalytic hydrodeoxygenation. Energy Environ Sci 7(1):103–129. https://doi.org/10.1039/C3EE43081B

Sankaranarayanan TM, Banu M, Pandurangan A, Sivasanker S (2011) Hydroprocessing of sunflower oil-gas oil blends over sulfided Ni-Mo-Al-zeolite beta composites. Bioresour Technol 102(22):10717–10723. https://doi.org/10.1016/j.biortech.2011.08.127

Satyarthi JK, Srinivas D (2011) Fourier transform infrared spectroscopic method for monitoring hydroprocessing of vegetable oils to produce hydrocarbon-based biofuel. Energy Fuels 25(7):3318–3322. https://doi.org/10.1021/ef200722q

Shams Z, Moghiman M (2018) Effect of metal oxide nanoparticles on the ignition characteristics of diesel fuel droplets: an experimental study. J Braz Soc Mech Sci Eng 40(2):1–10. https://doi.org/10.1007/s40430-018-1010-2

Singh D, Subramanian KA, Bal R, Singh SP, Badola R (2018) Combustion and emission characteristics of a light duty diesel engine fueled with hydro-processed renewable diesel. Energy 154:498–507. https://doi.org/10.1016/j.energy.2018.04.139

Somandepalli V, Kelly S, Davis S (2008) Hot surface ignition of ethanol-blended fuels and biodiesel. SAE Tech Pap 2008(724). https://doi.org/10.4271/2008-01-0402

Sonthalia A, Kumar N (2017) Hydroprocessed vegetable oil as a fuel for transportation sector: a review. J Energy Inst. https://doi.org/10.1016/j.joei.2017.10.008

Varuvel EG, Sonthalia A, Subramanian T, Aloui F (2018) NOx-smoke trade-off characteristics of minor vegetable oil blends synergy with oxygenate in a commercial CI engine. Environ Sci Pollut Res 25(35):35715–35724. https://doi.org/10.1007/s11356-018-3484-y

Vermeulen R, Silverman DT, Garshick E, Vlaanderen J, Portengen L, Steenland K (2014) Exposure-response estimates for diesel engine exhaust and lung cancer mortality based on data from three occupational cohorts. Environ Health Perspect 122(2):172–177. https://doi.org/10.1289/ehp.1306880

Zarchin R, Rabaev M, Vidruk-Nehemya R, Landau MV, Herskowitz M (2015) Hydroprocessing of soybean oil on nickel-phosphide supported catalysts. Fuel 139:684–691. https://doi.org/10.1016/j.fuel.2014.09.053

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

International Journal of Research in Engineering and Innovation (IJREI)
journal home page: http://www.ijrei.com
ISSN (Online): 2456-6934

Comparison of half effect absorption-compression cascaded refrigeration system using thermodynamic (energy-exergy) analysis

R.S. Mishra

Department of Mechanical & Production Engineering, Delhi Technological University Delhi, India

Abstract

In the most of industrial processes, the waste heat energy increases production cost. Using this waste heat to produce power or cooling can decrease the costs and increases the thermo-economic efficiency of the industrial systems. A growing need for refrigeration in industry due to demands of higher living standards, the increasing requirements for comfort, and the increasing thermal load a half effect, single & double effect and triple effect LiBr-H_2O vapour absorption refrigeration systems provide very good efficiency for cooling up to 0^0C temperature . The energy-exergy analysis in the vapour absorption and compression systems helped to find out the major loss incurring component. In order to reduce those losses many attempts has been made so that first law efficiency (COP) of the vapour absorption refrigeration system cascaded with vapour compression cycle can be increased. It is found that the thermodynamic performances in the case of cascaded half effect vapour absorption refrigeration system coupled with vapour compression cycle is improved by 44.65% increment of first law efficiency (i.e. over all COP), 172.867% increment of second law efficiency (i.e. exergetic efficiency) of the half effect vapour absorption refrigeration cascaded with vapour compression cycle using HFC-134a, 42.867% enhancement in first law efficiency (COP) of 142.7267% increment of second law efficiency using HFO -1234yf for -50ºC of evaporator temperature of VCRS . Similarly 72.024% reduction in exergy destruction ratio based on exergy of output of the half effect vapour absorption refrigeration cascaded with vapour compression cycle using HFC-134a and 70.44% reduction in exergy destruction ratio using HFO-1234yf ecofriendly refrigerant for -50ºC of evaporator temperature of VCRS. The performances of single effect cascaded vapour absorption refrigeration system coupled with vapour compression cycle significantly higher than cascaded half effect vapour absorption refrigeration coupled with vapour compression cycle.
© 2019 ijrei.com. All rights reserved

Keywords: Energy-exergy analysis, Absorption-compression cascaded refrigeration system, Thermodynamic analysis

1. Introduction

For the rapid growth and development of the countries there has been continues demand of energies. These energies are generally meet with burning of fossil fuels which has led to global warming. So for meeting the huge demand of Refrigeration and air conditioning at large scale Vapor absorption system working on Lithium Bromide and water is used. Solar energy and Waste heat from industries are the most favorable options because these sources do not degrade the environment anyway.

The research work done on VAS working on LiBr/H2O for cooling purpose is based on First and Second Law of thermodynamics has been done in context of one of its major component Generator. The researchers have tried to find out the optimum COP of the system varying the different parameters affecting its performance. They found that Generator temperature has highest influence on the COP of the system.

1.1 Energy-exergy analysis of vapour absorption system

Energy analysis of vapour absorption refrigeration system and its effect on Generator The basic thermodynamic analysis is the Energy analysis for any system. Energy analysis helps to find out the major loss incurring component. In order to reduce

Corresponding author: R. S. Mishra
Email Address: hod.mechanical.rsm@dtu.ac.in

those losses many attempts has been made so that COP of the cycle can be increased. Energy from external source is given in Generator so generator temperature is found to play an important role in COP of the cycle. The second laws, have been employed to be a tool for consideration of the quantity and quality of energy in the thermal system. Exergy analysis has determined the destruction and losses of exergy in various components and hence exergy destruction ratio which is calculated based on exergy of overall system input or exergy of overall system output. Refrigeration systems are generally driven by electrical power in the mechanical vapour compression system. Absorption chillers are an alternative approach; whose use the thermal power (industrial waste heat, renewable energy sources, or other thermal sources) as a driving force. Furthermore, absorption cooling systems do not cause ozone depletion, as long they use natural substances, as working fluids .A lot of researchers have studied various absorption refrigeration systems which use different thermal energy sources.

A parametric study of double effect series flow and double effect parallel flow vapour absorption machines is carried out by Meraj , Adnan Hafiz , M. Jamil Ahmad [1] to provide design engineers, appropriate information for developing sustainable machines to reduce irreversibility in each component of the whole system, thereby reducing net exergy loss and efforts have been made by performed Second law analysis and found that the exergetic efficiency first increases , reaches a maxima and then starts decreasing with increase in high pressure generator temperature and also concluded that the exergetic efficiency is more in parallel flow as compared to the series flow for the same generator temperatures in the vapour absorption refrigeration system. Similarly irreversibility first decreases in both cases, reaches a minimum and then increases. Irreversibility is less in parallel configuration as compared to the series one. Absorption cooling cycles are environmental and can use solar or waste heat for cooling with very small electric power.

SC Kaushik, Akhilesh Arora,[1] carried out the energy and exergy analysis of single effect and series flow double effect water–lithium bromide absorption systems by developing computational model using water–lithium bromide solution. This analysis involved the determination of effects of generator, absorber and evaporator temperatures on the energetic and exergetic performance of these absorption refrigeration systems. The effects of pressure drop between evaporator and absorber, and effectiveness of heat exchangers are also investigated for vapour absorption systems. The numerical computation was carried out for finding performance parameters (i.e. coefficient of performance, exergy destruction, efficiency defects and exergetic efficiency) and it is found that the coefficient of performance of the single effect system lies in range of 0.6–0.75 and 0.90 to 1.2 for double effect system respectively. Mohd. Meraj et.al [2] studied double effect series flow and double effect parallel flow vapour absorption machines and carried out second law analysis to reduce irreversibility in each component of the whole system, thereby reducing net exergy loss and found that the exergetic efficiency first increases , reaches a maxima and then starts decreasing with increase in high pressure generator temperature. . It has also been observed that the exergetic efficiency is more in parallel flow as compared to the series flow for the same generator temperatures Rabi Karaali [3] carried out work on thermal analysis of a double effect parallel flow and single effect absorption cooling systems by developing computer program for the thermodynamic properties of lithium bromide-water solutions using the exergy analysis and found that the double effect parallel flow absorption systems have better advantages than the single effect absorption system. The coefficient performance (COP) and the exergetic efficiency (exergetic coefficient of performance) of the double effect parallel flow absorption systems (i.e. the double effect cycle COP is 1.195 and exergetic efficiency as 0.28) , are higher than the single effect cycle (COP is 0.68) and exergetic efficiency (as 0.23) . Rabi Karaali [4] presented exergy analysis of a half effect absorption system using thermodynamic properties of lithium bromide-water solutions, by developing computer program in FORTRAN codes. Due to the reason that the half effect absorption systems is used for low temperature heat energy or waste heat energy and found that the coefficient performance (COP) and the exergetic efficiency of the half effect absorption system are 0.45 and 0.24, respectively. For each component the exergy loss is calculated and it was found that the 92 % of total irreversibilities in the absorber and in the evaporator and also concluded that better design and efficiencies of the evaporator and the absorber is very important for the half effect absorption cycles. By improving the design of these two components will directly improve and affect the working conditions and thermal performances of the cycle. Md Azhar and M. Altamush Siddiqui [5] analyzed Triple Effect Vapour Absorption cycle using liquefied petroleum gas (LPG) and compressed natural gas (CNG) as sources of energy selected because the triple effect system requires heat at relatively high temperature using LiBr-H2O solution as working pair and compared with the Single and Double Effect cycles. The thermodynamic analysis is done for different values of the evaporator and the condenser temperatures; the absorber temperature is assumed to be equal to the main condenser temperature by assuming the temperature and concentration of LiBr salt in the high pressure generator, to which heat is supplied, are varied simultaneously for fixed temperatures/pressures in the middle and high pressure condensers and investigated the effect of the high pressure generator temperature on the coefficient of performance (COP), volume flow rate and cost of the LPG and CNG. Arora et al. [6] have developed a double effect cycle 120-150^{0}C, half effect 60-75^{0}C and triple effect cycles which require generator temperature even higher than 150^{0}C.

Dixit Manoj et.al [7] developed mathematical model of vapour absorption-compression cascaded refrigeration system, comprising of a vapour compression refrigeration system in low temperature stage using CO2, NH3 and R134a as ecofriendly refrigerants and H_2O-LiBr vapour absorption

refrigeration system at the high temperature stage and found that the cascade condenser, compressor and refrigerant throttle valve are the major source of exergy destruction**.** S Arivazhagan et.al. [8] carried out experimental investigation on the performance of a two-stage half effect vapour absorption cooling system The prototype is designed for 1 kW cooling capacity using HFC based working fluids (R134a as refrigerant and DMAC as absorbent). The performance of the system in terms of degassing range, coefficient of performance and second law efficiency has been obtained. The system is capable of producing evaporating temperature as low as -7^0C with generator temperatures ranging from 55^0 to 75^0C. The degassing range is 40% more in high absorber than in low absorber as the high absorber is operated at optimum intermediate pressure. The optimum generator temperature is found to be in the range of 65–70^0C at which the coefficient of performance is 0.36. Shun-Fu Lee and S. A. Sheriff [9] carried out parametric analysis using thermodynamic analysis of a lithium bromide/water absorption system for cooling and heating applications and it was found that a low cooling water temperature yields both a higher cooling COP and higher exergetic efficiency and also observed that by increasing the heat source generator temperature can improve the COP of the system and also found that the heat source temperature increases beyond a certain limit , the COP of the system decreases due to the negative effect of increasing the heat source temperature. Inzunza et al. [10], compared of the performance of single-effect, half-effect, double-effect in series and inverse absorption cooling systems operating with the mixture LiBr- H_2O and found that for the generation temperature between 100^0C and 110^0C, the COP of the single effect was up to 0.89, for the generation temperature of over 55 ^{0}C the COP of the half effect was up to 0.44. They also found that the most efficient system is double effect of the COP up to 1.48 for higher temperature while, for low temperatures the half effect systems work better than any other. Wang and Zheng [11] developed the performance of half effect & single effect LiBr / H_2O absorption cooling cycle and found that half effect cycle's absorption cooling cycle can be used for generator temperature from 50^0C to 80^0C. This paper mainly deals with half effect vapour absorption-compression cascade system is lowest by using R141b and highest while using R227ea as shown in table-2. The cascade system exergetic refrigeration system, comprising of a vapour compression cascade refrigeration system in low temperature stage and half effect absorption refrigeration system at the high temperature stage, is analyzed. CO_2, NH3 and R134a have been considered as refrigerants in the compression stage and the H2O-LiBr refrigerant absorbent pair in the absorption stage using energy-exergy analysis. The performance of the refrigeration system. The comparative study carried out in terms of their thermal performances using HFC-134a and HFo-1234yf for evaporator temperature of VCRS is -50^0C and -30^0C respectively using HFO-1234ze respectively.

2. Results and Discussion

Following numerical values have been taken for validation of thermal models for half effect LiBr-H_2O cascade vapour compression-absorption refrigeration systems.

2.1 Half Effect Cascade vapour compression-absorption refrigeration system using ecofriendly refrigerants

$T_{_Generator}$=80ºC.T_Cond=30ºC, $T_{_Eva}$=5ºC, T_absorber=30ºC, Compressor Efficiency=0.80, T_Eva_LTC=-50ºC (for using R-1234yf and HFCs, hydrocarbons, natural refrigerants in VCRS), & -30ºC (for using R1234ze & R1234yf, HFCs, hydrocarbons, natural refrigerants in the VCRS), approach (Temperature Overlapping) =10ºC)

Table-1 shows the variation of thermal performance parameters of triple effect cascade vapour absorption refrigeration system using lithium bromide -water solution coupled with vapour compression refrigeration system using 29.09 kW of load. The evaporator of VARS is cascaded with condenser of vapour compression refrigeration systems using following nine ecofriendly refrigerants. It is found that overall first law efficiency in terms of coefficient of performance (COP) of cascade refrigeration system using R245fa refrigerant is highest and lowest by using R227ea as shown inTable-1(a) and Table-1(b) respectively while exergy destruction of
efficiency is highest by using R141b and lowest by using R227ea as shown in table-3.

Table-1(a):Thermal performance of single Effect LiBrH_2O Vapour absorption refrigeration system cascaded by vapour compression refrigeration system using ecofriendly refrigerants at the evaporator temperature of -30ºC in the VCRS cycle

S.No	Refrigerant	$COP_{_Over_alll}$	$EDR_{_Overall}$	Exergetic efficiency
1	R134a	1.126	2.057	0.3271
2	R1234ze	1.124	2.070	0.3257
3	R1234yf	1.117	2.122	0.3203
4	R227ea	1.117	2.218	0.3107
5	R236fa	1.119	2.108	0.3217
6	R245fa	1.133	2.008	0.3324
7	R32	1.122	2.087	0.3239
8	R507a	1.110	2.174	0.3151
9	R717	1.129	2.04	0.3289

10	R123	1.138	1.976	0.3360
11	R125	1.099	2.260	0.3068
12	R152a	1.134	2.003	0.3330
13	R404a	1.105	2.216	0.3109
14	R410a	1.122	2.086	0.3240
15	R407c	1.004	2.382	0.2957
16	R290	1.125	2.069	0.3258
17	R600	1.133	2.011	0.3321
18	R600a	1.125	2.063	0.3265
19	R141b	1.145	1.929	0.3414
20	R143a	1.141	2.154	0.3171

Table-1(b)::Thermal performance of Half Effect LiBrVapour absorption refrigeration system cascaded by vapour compression refrigeration system using ecofriendly refrigerants (for T_EVA_VCRS= -50°C)

S.No	Refrigerant	$COP_{_Over_alll}$	$EDR_{_Overall}$	Exergetic Efficiency
1	R134a	0.6155	1.516	0.3974
2	R1234yf	0.6079	1.602	0.3844
3	R227ea	0.5956	1.746	0.3642
4	R236fa	0.6084	1.596	0.3853
5	R245fa	0.6196	1.473	0.4044
6	R32	0.6138	1.535	0.3945
7	R507a	0.6042	1.644	0.3782
8	R717	0.6157	1.515	0.3977
9	R123	0.6238	1.428	0.4118
10	R125	0.5951	1.752	0.3634
11	R404a	0.6011	1.680	0.3732
12	R410a	0.6142	1.532	0.3950
13	R407c	0.5956	1.746	0.3641
14	R290	0.6151	1.521	0.3967
15	R600	0.6203	1.465	0.4057
16	R600a	0.6144	1.528	0.3955
17	R141b	0.6296	1.368	0.4223
18	R143a	0.6068	1.614	0.3825

Table-2:Thermal performance of double Effect LiBrH_2O Vapour absorption refrigeration system cascaded by vapour compression refrigeration system using ecofriendly refrigerants at the evaporator temperature of -30°C in the VCRS cycle,

S.No	Refrigerant	$COP_{_Over_alll}$	$EDR_{_Overall}$	Exergetic Efficiency
1	R134a	1.822	1.037	0.4910
2	R1234ze	1.817	1.047	0.4884
3	R1234yf	1.80	1.085	0.4796
4	R227ea	1.77	1.156	0.4638
5	R236fa	1.805	1.075	0.4820
6	R245fa	1.838	1.001	0.4998
7	R32	1.813	1.057	0.4863
8	R507a	1.784	1.123	0.4711
9	R717	1.829	1.021	0.4949
10	R123	1.849	0.9767	0.5059
11	R125	1.757	1.187	0.4573
12	R152a	1.84	0.9964	0.5009
13	R404a	1.772	1.152	0.4647
14	R410a	1.812	1.059	0.4858
15	R407c	1.731	1.251	0.4442
16	R290	1.818	1.045	0.4889
17	R600	1.837	1.003	0.4994
18	R600a	1.819	1.041	0.4899
19	R141b	1.865	0.9424	0.5148
20	R143a	1.79	1.108	0.4744

Table-3:Thermal performance of Tripple Effect $LiBrH_2O$ Vapour absorption refrigeration system cascaded by vapour compression refrigeration system using ecofriendly refrigerants at the evaporator temperature of -30ºC in the VCRS cycle,

S.No,1	Refrigerant	$COP_{_Over_alll}$	$EDR_{_Overall}$	Exergetic Efficiency
1	R134a	2.151	1.421	0.4131
2	R1234ze	2.148	1.429	0.4116
3	R1234yf	2.131	1.469	0.4050
4	R227ea	2.10	1.543	0.3933
5	R236fa	2.135	1.459	0.4067
6	R245fa	2.168	1.382	0.4198
7	R32	2.139	1.448	0.4084
8	R507a	2.113	1.510	0.3984
9	R717	2.154	1.414	0.4142
10	R123	2.179	1.357	0.4242
11	R125	2.087	1.574	0.3885
12	R152a	2.169	1.380	0.4202
13	R404a	2.098	1.547	0.3926
14	R410a	2.142	1.442	0.4095
15	R407c	2.034	1.712	0.3687
16	R744	1.993	1.827	0.3537
17	R290	2.147	1.430	0.4115
18	R600	2.167	1.385	0.4194
19	R600a	2.148	1.425	0.4124
20	R141b	2.196	1.321	0.4309
21	R143a	2.12	1.495	0.4008

Table 4(a): Performance comparison of single effect vapour absorption refrigerated cascaded with vapour compression refrigeration system using HFC-134a at evaporator temperature of -50ºC .

Performance Parameters	Half effect Vapour Absorption System	Half effect Cascade VARS_VCRS	% improvement
COP	0.4255	0.6155	44.65
Exergetic Efficiency	0.1558	0.3974	155
EDR	5.419	1.516	72.022 (Reduction)

The performance parameters have been compared with the half effect vapour absorption system with half effect vapour absorption-compression refrigeration system cascaded with vapour compression refrigeration system using ecofriendly HFC-134a and it was observed that 44.65% first law efficiency enhancement and 42.876% using HFO-1234yf for -50ºC of evaporator temperature of vapour compression cycle as shown in table-4(a) & table-4(b) respectively. Similarly 155% enhancement in second law efficiency using half effect vapour absorption system cascaded with vapour compression refrigeration cycle using HFC-134a and 146.726% using HFO-1234yf in the vapour compression cycle as shown in table-4(a-b) respectively. Similarly exergy reduction is 72.024% using HFC-134a and 146.726% using HFO-1234yf in the vapour compression cycle as shown in table-4(a) & table-4(b) respectively. The performance parameters have been compared with the half effect vapour absorption system with half effect vapour absorption-compression refrigeration system cascaded with vapour compression refrigeration system using ecofriendly HFC-134a, single effect vapour absorption system with half effect vapour absorption-compression refrigeration system cascaded with vapour compression refrigeration system using ecofriendly HFC-134a, double effect vapour absorption system with half effect vapour absorption-compression refrigeration system cascaded with vapour compression refrigeration system using ecofriendly HFC-134a, tripplel effect vapour absorption system with half effect vapour absorption-compression refrigeration system cascaded with vapour compression refrigeration system using ecofriendly HFC-134a respectively and it is found that first law efficiency (COP) and second law efficiency (exergetic efficiency) performance increases while exergy destruction based on exergy of product is decreases

.

Table 4(b) : Performance comparison of single effect vapour absorption refrigerated cascaded with vapour compression refrigeration system using HFO-1234yf at evaporator temperature of -50ºC .

Performance Parameters	Half effect Vapour Absorption System	Half effect Cascade VARS_VCRS	% improvement
COP	0.4255	0.6079	42.876
Exergetic Efficiency	0.1558	0.3844	146.726
EDR	5.419	1.602	70.44

Table-5 : Performance comparison of single effect vapour absorption refrigerated cascaded with vapour compression refrigeration system using HFO-1234yf at evaporator temperature of -50°C .

Performance Parameters	Half effect simple vapour absorptin refrigeration system	Half effect cascaded vapour Absorption System	Single effect Cascaded vapour absorption System	Double Effect cascaded vapour absorption System	Triple effect cascaded vapour absorption System
COP	0.4255	0.6155	1.126	1.822	2.151
Exergetic Efficiency	0.1558	0.3974	0.3271	0.4131	0.4910
EDR	5.519	2.057	1.516	1.421	1.037

3. Conclusions

The following conclusions were drawn.

(i) The thermodynamic performances in the case of cascaded half effect vapour absorption refrigeration system coupled with vapour compression cycle is improved.

(ii) The performances of single effect cascaded vapour absorption refrigeration system coupled with vapour compression cycle significantly higher than cascaded half effect vapour absorption refrigeration coupled with vapour compression cycle.

(iii) The 44.65% increment of first law efficiency (i.e. over all COP) of the half effect vapour absorption refrigeration cascaded with vapour compression cycle using HFC-134a ecofriendly refrigerant for -50°C of evaporator temperature of VCRS.

(iv) The 42.867% increment of first law efficiency (i.e. over all COP) of the half effect vapour absorption refrigeration cascaded with vapour compression cycle using HFC-134a ecofriendly refrigerant for -50°C of evaporator temperature of VCRS.

(v) The 172.867% increment of second law efficiency (i.e. exergetic efficiency) of the half effect vapour absorption refrigeration cascaded with vapour compression cycle using HFC-134a ecofriendly refrigerant for -50°C of evaporator temperature of VCRS.

(vi) The 142.7267% increment of second law efficiency (i.e. exergetic efficiency) of the half effect vapour absorption refrigeration cascaded with vapour compression cycle using HFC-134a ecofriendly refrigerant for -50°C of evaporator temperature of VCRS.

(vii) The 72.024% reduction in exergy destruction ratio based on exergy of output of the half effect vapour absorption refrigeration cascaded with vapour compression cycle using HFC-134a ecofriendly refrigerant for -50°C of evaporator temperature of VCRS.

(viii) The 70.44% reduction in exergy destruction ratio based on exergy of output of the half effect vapour absorption refrigeration cascaded with vapour compression cycle using HFC-134a ecofriendly refrigerant for -50°C of evaporator temperature of VCRS.

References

[1] SC Kaushik, A Arora [2009] Energy and exergy analysis of single effect and series flow double effect water–lithium bromide absorption refrigeration systems, International journal of Refrigeration, Vol-32, No-6page-1247-1258

[2] Mohd. Meraj , Adnan Hafiz , M. Jamil Ahmad [2016] Second Law Analysis of Series and Parallel Flow Double Effect Vapor Absorption Chiller, IOSR Journal of Mechanical and Civil Engineering (IOSR-JMCE) e-ISSN: 2278-1684,p-ISSN: 2320-334X, Volume 13, Issue 3 Ver. VI (May- Jun. 2016), PP 104-110.

[3] Rabi Karaali [2017] Exergetic Comparison of Single and Double Effect Absorption Cooling Cycles,Karaelmas Fen ve Müh. Derg. 7(1):122-128, 2017.

[4] Rabi Karaali [2016] Exergy analysis of half-effect absorption cooling cycles, international journal of engineering sciences & research technology (IJESRT).ISSN: 2277-9655,Vol-5, issue-9 .page- 405-411.

[5] Md Azhar and M. Altamush Siddiqui[2013] Thermodynamic analysis of a gas operated triple effect absorption cycle, International Journal of Innovative Research in Science, Engineering and Technology , Vol. 2, ISSN: 2319-8753 , Issue 5.

[6] Akhilesh Arora,Manoj Dixit,S.C. Kaushik.[2018] : Computation of optimum parameters of a half effect water-lithium bromide vapour absorption refrigeration system, Journal of Thermal Engineering, Vol-2, No-2, page- 683-692.

[7] M Dixit, SC Kaushik, A Arora [2018] Energy and exergy analysis of absorption-compression cascade refrigeration system, Journal of Thermal Engineering , Vol-2 (6), page-995-1006.

[8] Arivazhagan S, et.al. [2006]Experimental studies on HFC based two-stage half effect vapour absorption cooling system, Applied Thermal Engineering, Vol-26, Issues-14-15, Pages 1455-1462.

[9] Lee Shun.Fu., Sherif, S.A. [2000] Thermodynamic analysis of a lithium bromide/water absorption system for cooling and heating applications, International Journal of energy Research,Vol.25, pp.1019-1031.

[10] L.A. Domínguez-Inzunza, J.A. Hernández-Magallanes, M. Sandoval-Reyes, and W. Rivera[2014], "Comparison of the performance of single-effect, half-effect, double-effect in series and inverse and triple-effect absorption cooling systems operating with the NH3-LiNO3 mixture" Applied Thermal Engineering VOL.66, pp.612- 620,.

[11] J, Wang and D. Zheng, [2009]"Performance of one and half effect absorption cooling cycle of H_2O/LiBr system", Energy Conversion and Management, VOL. 50, pp. 3087-3095.

Cite this article as: R.S. Mishra, Comparison of half effect absorption-compression cascaded refrigeration system using thermodynamic (energy-exergy) analysis, , *International journal of research in engineering and innovation (IJREI)*, vol 3, issue 1 (2019), 6-11.

IOP Conference Series: Materials Science and Engineering

PAPER • **OPEN ACCESS**

Comparison of properties of ternary fuel blends of diesel-octanol with biodiesel

To cite this article: Sidharth and Naveen Kumar 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **691** 012050

View the article online for updates and enhancements.

This content was downloaded from IP address 14.139.251.100 on 19/12/2019 at 04:50

Comparison of properties of ternary fuel blends of diesel-octanol with biodiesel

Sidharth[1,2], **Naveen Kumar**[2]

[1]Mechanical and Automation Engineering Department, Maharaja Agrasen Institute of Technology, Delhi, India
[2]Centre for Advanced Studies and Research in Automotive Engineering, Delhi Technological University, Delhi, India

Email: sidraje2002@yahoo.com

Abstract. Biodiesel is adopted by many countries as a substitute to diesel and is used in blended form in a diesel engine. It has very similar properties to diesel. However, owing to slightly higher viscosity, density and nitrogen oxide (NO_X) emissions on burning, it still is not the best choice. Octanol, have similar properties to diesel. Octanol has added oxygen and also reduces the NO_X emissions due to the quenching effect in the cylinder. However, its lower calorific value shall reduce engine performance. In the current study, waste cooking oil biodiesel is produced by a single step transesterification process and several blends of diesel, biodiesel and octanol were prepared. Fourier-transform infrared spectroscopy (FTIR) and oxidation stability tests of the produced biodiesel were performed. FTIR test confirmed the presence of biodiesel. Oxidation stability test of biodiesel was also passed as the induction time in over 6 hours. Blends of diesel and octanol were prepared in proportion of diesel (95%, 90%, 85% and 80%) and octanol (5%, 10%, 15% and 20%). These blends were added to 3%, 5% and 10% biodiesel. Several physico-chemical properties namely density, viscosity, calorific value, flash point and cold filter plugging point (CFPP) were determined as per ASTM standards. It is concluded that the 10% octanol, 90% diesel when mixed with 10% biodiesel showed very similar properties to diesel and shall be a viable alternative to petroleum diesel.

Keywords: octanol, diesel, biodiesel, density, viscosity, calorific value

1. Introduction

Diesel engines are the backbone of the Indian economy because of their major usage in transportation, agriculture and power sectors. India is facing the problem of pollution due to petroleum diesel and its limited resources. This has led the researchers to seek an alternative for diesel. Vegetable oils have been used at the beginning of the 20th century by Rudolf Diesel [1] and quoted as a feasible fuel for the diesel engine. However, many researches have been carried in the last few decades which have proven the usage of vegetable oil as an undesirable choice in the current high-speed engines. High viscosity, higher engine deposits, injector coking, piston ring sticking [2,3] are the major problems of using them in a diesel engine. Vegetable oils also have higher engine emissions and lower performance characteristics as compared to diesel [4–6]. The researchers are giving serious attention to biodiesel as a replacement to diesel owing to its similar properties [7–9] which may help to cope with the depleting

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

petroleum reserves. A long chain alkyl ester named as biodiesel is a very promising eco-friendly fuel compared to vegetable oil because of its lower viscosity, fewer engine deposits, higher cetane number, better performance and emission characteristics [10–13]. Therefore biodiesel is a viable option for developing countries who aim to be self-sustainable and also for saving the foreign exchange in procuring crude oil from oil-rich countries.

Octanol is another promising fuel to be used in the diesel engine due to its higher cetane number than other low carbon chain alcohols. Moreover, added oxygen in octanol [14] shall also help in better combustion and lower emissions. However, octanol cannot be used directly in diesel engine due to its lower heating value and higher viscosity. It is, therefore, preferable to use octanol partially in diesel engine in the form of blends. Table 1 summarizes the works carried in the past with ternary fuel blends for comparing their physico-chemical properties.

Table 1. Effect of adding different ternary fuel blends on the properties of diesel

References	**Ternary Fuels**	**Effect of adding different fuels in diesel on physico-chemical properties**
Atmali[15]	• Diesel • Biodiesel • Propanol /n-butanol/1-pentanol	• Density, kinematic viscosity, flash point and CFPP increase on increasing biodiesel content and decrease on increasing the alcohol content. • Lower heating value and Cetane number decrease in increasing biodiesel and alcohol percentage.
Redel-Macías*et al.*[16]	• Diesel • Ethanol • Castor oil	• CFPP increase on increasing oil percentage and decrease on increasing ethanol percentage. • Viscosity increase on increasing oil percentage and decrease on increasing ethanol percentage. • Calorific value decrease on increasing oil or ethanol percentage.
EL-Seesy*et al.*[17]	• Diesel • Ethanol • Castor oil	• Density increase on increasing oil content and decrease on increasing ethanol content. • Viscosity increase on increasing oil percentage and decrease on increasing ethanol percentage. • Calorific value decrease on increasing oil or ethanol percentage Calculated cetane index decrease on increasing oil or ethanol percentage.
Nanthagopal*e t al.* [18]	• Diesel • Decanol • Calophyl lumInophyl lum biodiesel	• Calorific value and cetane index decrease in increasing biodiesel or decanol percentage. However, diesel and decanol have nearly equal calorific values. • Density, flash point and kinematic viscosity increase on increasing decanol and biodiesel percentage.
Babu D *et al.* [19]	• Diesel • Biodiesel • n-pentanol/n-hexanol	• Density, flash point, cloud point and pour point increase on increasing biodiesel percentage and decrease on increasing n-pentanol/n-hexanol percentage. • Viscosity increase in increasing biodiesel and n-pentanol/n-hexanol content. • Calorific value decrease in increasing biodiesel and n-pentanol/n-hexanol.
Qi *et al.*[20]	• Diesel • Biodiesel • Methano l	• Density and viscosity increase on increasing biodiesel percentage and decrease on increasing methanol percentage. • Calorific value decrease in increasing biodiesel and methanol content. • Oxygen content increase with both biodiesel and methanol.

In the current study, several blends of diesel, waste cooking oil biodiesel (WCOB) and octanol were formulated and their properties were determined. The physico-chemical properties of waste cooking oil biodiesel were also determined for this study. Oxidation stability and Fourier-transform infrared spectroscopy (FTIR) tests were also carried out for the WCO biodiesel. Waste cooking oil was for the present work owing to the concern that most of the restaurants and cafeterias dispose of the waste oil in the drains as per the practice employed in the industry. This oil shall mix with the water bodies and contaminate it. Therefore, the oil can be used to produce biodiesel due to the favorable properties.

2. Material and Methods

Waste cooking oil was procured from a local restaurant in New Delhi, India. Chemicals of analytical grade were purchased from a local vendor. Free fatty acid (FFA) of the oil is determined using equation 1 and 2.

$$\text{FFA}=\frac{Acid\ Number}{2} \tag{1}$$

$$\text{Acid number} = \frac{\left(56.1\times amount\ of\ 0.1N\ sodium\ hydroxide\ solution\ consumed\ \right)}{Weight\ of\ oil\ sample\ taken} \tag{2}$$

Biodiesel is mainly produced from vegetable oil or animal fats by following both the esterification and transesterification processes or only by the transesterification process[21]. The process to be followed depends upon the free fatty acid (FFA) content of the fuel[22]. If the FFA of the fuel is more than 2 then a two-step process (both esterification and transesterification) is to be carried out for conversion of vegetable oil or animal fat to biodiesel. If the FFA is less than 2 then a direct transesterification process can be carried out. In the current study, waste cooking oil was used to make methyl ester in the presence of alcohol (methanol) and base catalyst (potassium hydroxide) through transesterification process.

For producing biodiesel, WCO was first passed through a fine sieve to remove heavy particles and then it is passed through filter papers to remove unwanted particles. The oil is then heated over a hot plate to reach a temperature of over 100°C to remove moisture from it. The oil is kept at that temperature for nearly 30 minutes. The oil is then cooled to the desired temperature. Catalyst (potassium hydroxide- 0.05% weight of oil) is first mixed thoroughly with methanol (20% volume of oil). The mixture of catalyst and oil is then added to the oil. This mixture of oil, catalyst and methanol is then constantly stirred over the hot plate for the desired time. The methyl ester or biodiesel so formed is allowed to settle in the separating funnel to separate out the heavy glycerine. Figure 1 shows the separation of Glycerine and biodiesel in a separating funnel.



Figure 1. Separation of Glycerin (bottom) and Biodiesel (Top)

Excess methanol is collected by heating the biodiesel. Water washing was also carried out for removing any other un-dissolved chemicals from the biodiesel and then heating it to above 100°C to remove water from it.
Octanol having 99% purity is mixed in diesel in the proportions 95:5, 90:10, 85:15 and 80:20 in percentage volume basis. Varying amount of biodiesel 3%, 5% and 10% were added to the above-mentioned proportions of the test fuels to form blends of ternary fuel.
Table 2 shows different properties of diesel, biodiesel and Octanol determined using different setups. The equipments used for the measurement of density- Anton Parr DMA 4500, Viscosity-Capillary tube, Calorific value- Parr 6100 oxygen bomb calorimeter, Flash point- BavenoPensky Martens apparatus, Cold filter plugging point (CFPP)- Normalab automatic NTE 450, Oxidation stability- Metrohm 873 Rancimat and FTIR- Thermo scientific Nicolet 380.

3. Result and Discussion

The test fuels of diesel, octanol and WCO biodiesel were kept in an undisturbed state for over 4 months. No phase separation is observed in any of the test fuel blends. FTIR test results and oxidation stability results of the produced biodiesel are shown in Figure 2 and Figure 3.

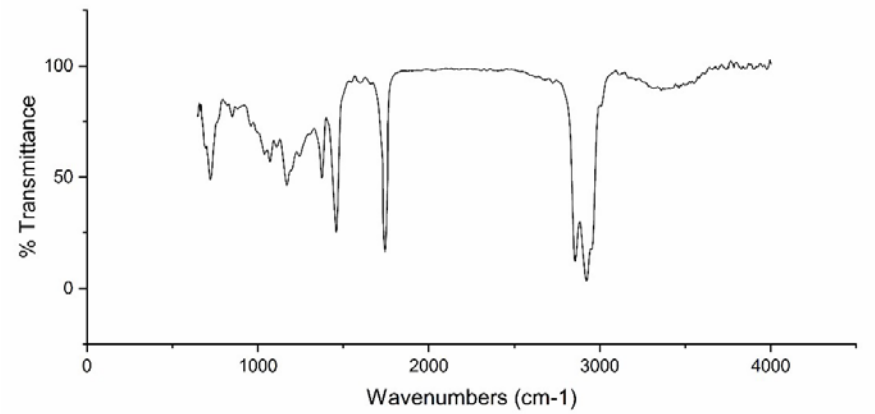


Figure 2.FTIR of Waste cooking oil biodiesel

It is seen from Figure 2 that a large peak at 1750 cm^{-1} shows the presence of biodiesel. This peak is very evident as the transesterification reaction sometimes produce the compounds which are not methyl esters and they may behave in a different manner in a diesel engine. Therefore the presence of this peak confirms the transesterification process is successful to produce biodiesel.

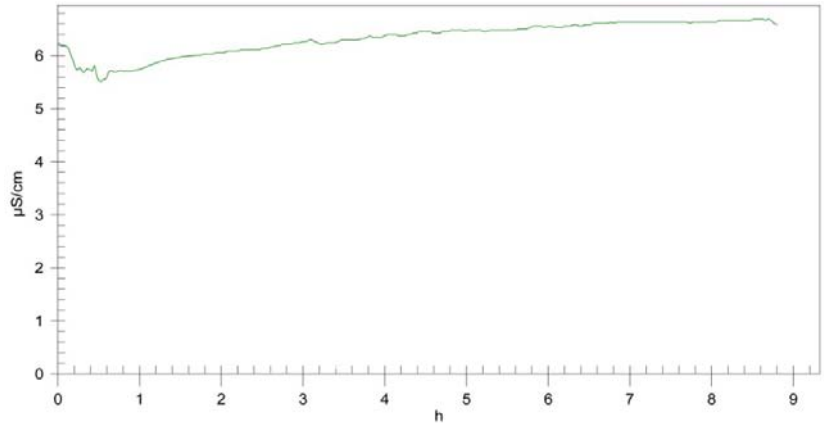


Figure 3. Oxidation stability test of WCO biodiesel

The oxidation stability gives us an idea about the presence of unsaturated fats in the methyl ester that may degrade the methyl ester. Since the test runs successfully for over 6 hours, therefore, the fuel qualifies both EN 14214 and ASTM D6751 standards.

It is seen from Table 2, that biodiesel has very high viscosity and density as compared to diesel. Moreover, octanol also has high viscosity as compared to petroleum diesel. It is therefore not advisable to use biodiesel and octanol in neat form. However, it is seen from the literature that oxygen content of biodiesel and octanol shall promote better combustion and hence produce less harmful emissions.

Therefore ternary fuel blends of the three are formed in different proportions and their properties are seen in Table 3 to Table 6. As discussed earlier, diesel and octanol are added in the proportions 95:5, 90:10, 85:15 and 80:20 in percentage volume basis. Varying amount of biodiesel 3%, 5% and 10% were added to the above-mentioned proportions of the test fuels to form blends of ternary fuel

Table 2. Properties of diesel, WCO biodiesel and octanol

S.No.	Property	Diesel	Biodiesel	Octanol
1	Density (at 15°C) in Kg/m^3	837.1	890.7	822.9
2	Viscosity (at 40°C) in cSt	2.51	5.91	4.98
3	Calorific value (MJ/Kg)	45.42	41.85	39.74
4	FFA	-	1.9	-
5	Flash Point (°C)	59	165	82
6	CFPP (°C)	-11	-6.1	-12.3

Table 3. Properties of 95% diesel and 5% octanol when blended with 3%, 5% and 10% WCO biodiesel

S.No.	Property	Diesel and octanol in 95:5 % volume		
		WCOB 3%	WCOB 5%	WCOB 10%
1	Density (at 15°C) in Kg/m^3	838.1	840.4	842.1
2	Viscosity (at 40°C) in cSt	2.74	2.81	2.98
3	Calorific value (MJ/Kg)	44.85	44.74	43.91
4	Flash Point (°C)	64	65	70
5	CFPP (°C)	-10.7	-10.4	-10.1

Table 4. Properties of 90% diesel and 10% octanol when blended with 3%, 5% and 10% WCO biodiesel

S.No.	Property	Diesel and octanol in 90:10 % volume		
		WCOB 3%	WCOB 5%	WCOB 10%
1	Density (at 15°C) in Kg/m^3	837.4	838.3	840.1
2	Viscosity (at 40°C) in cSt	2.89	2.93	3.09
3	Calorific value (MJ/Kg)	44.71	44.67	44.52
4	Flash Point (°C)	65	66	71
5	CFPP (°C)	-10.9	-10.7	-10.4

Table 5. Properties of 85% diesel and 15% octanol when blended with 3%, 5% and 10% WCO biodiesel

S.No.	Property	Diesel and octanol in85:15 % volume		
		WCOB 3%	WCOB 5%	WCOB 10%
1	Density (at 15°C) in Kg/m^3	837.2	837.5	841.8
2	Viscosity (at 40°C) in cSt	3.10	3.13	3.31
3	Calorific value (MJ/Kg)	44.44	44.38	44.15
4	Flash Point (°C)	66	68	73
5	CFPP (°C)	-11.0	-10.8	-10.4

Table 6. Properties of 80% diesel and 20% octanol when blended with 3%, 5% and 10% WCO biodiesel

S.No.	Property	Diesel and octanol in 80:20 % volume		
		WCOB 3%	WCOB 5%	WCOB 10%
1	Density (at 15°C) in Kg/m^3	836.6	837.1	839.9
2	Viscosity (at 40°C) in cSt	3.22	3.35	3.48
3	Calorific value (MJ/Kg)	44.16	44.02	43.81
4	Flash Point (°C)	67	68	74
5	CFPP (°C)	-11.1	-10.9	-10.6

It is seen from Table 3 to Table 6 that all the properties of the ternary fuel blends are very much similar to petroleum diesel. However, for more than 15% octanol blends viscosity tends to increase to a larger value. However, not much prominent change is observed for other properties of test fuel blends. Therefore, it is advisable to use to the ternary fuel blends in a diesel engine 10% octanol, 90% diesel mixed with 10% biodiesel.

4. Conclusions

The current research examined the properties of diesel, biodiesel and octanol. Biodiesel was prepared from Waste cooking oil using direct transesterification process owing to less than 2% FFA of oil. FTIR test was performed for the WCO biodiesel to verify the presence of methyl ester. A strong peak at 1750 cm^{-1} confirms for the same. Oxidation stability was also passed for the WCO biodiesel. Several properties namely density, viscosity, calorific value, flash point and CFPP were determined. The results indicated that when 10% octanol and 90% diesel is mixed with 10% biodiesel, a very minor change is observed in the properties of the ternary fuel blends. However, for blends containing 15% and 20% octanol, higher viscosities are observed. This may lead to problems in the engine in the long run. Therefore, ternary fuels blend of 10% octanol, 90% diesel when mixed with 10% biodiesel and shall be a possible substitute to current petroleum diesel. Performance, emission and combustion studies need to be carried out in the future to get a clear insight about the test fuel blends in a diesel engine.

5. References

[1] Dash S K and Lingfa P 2018 An overview of biodiesel production and its utilization in diesel engines *IOP Conf. Ser. Mater. Sci. Eng.***377** 012006

[2] Tarabet L, Loubar K, Lounici M S, Hanchi S and Tazerout M 2012 Eucalyptus Biodiesel as an Alternative to Diesel Fuel: Preparation and Tests on DI Diesel Engine *J. Biomed. Biotechnol.***2012** 1–8

[3] Pipitone E and Costanza A 2018 An experimental investigation on the long-term compatibility of preheated crude palm oil in a large compression ignition diesel engine *Biofuel Res. J.***5** 900–8

[4] Dabi M and Saha U K 2019 Application potential of vegetable oils as alternative to diesel

fuels in compression ignition engines: A review *J. Energy Inst.*
[5] Che Mat S, Idroas M Y, Hamid M F and Zainal Z A 2018 Performance and emissions of straight vegetable oils and its blends as a fuel in diesel engine: A review *Renew. Sustain. Energy Rev.***82** 808–23
[6] Gad M S, El-Araby R, Abed K A, El-Ibiari N N, El Morsi A K and El-Diwani G I 2018 Performance and emissions characteristics of C.I. engine fueled with palm oil/palm oil methyl ester blended with diesel fuel *Egypt. J. Pet.***27** 215–9
[7] Abed K A, El Morsi A K, Sayed M M, Shaib A A El and Gad M S 2018 Effect of waste cooking-oil biodiesel on performance and exhaust emissions of a diesel engine *Egypt. J. Pet.***27** 985–9
[8] Abed K A, Gad M S, El Morsi A K, Sayed M M and Elyazeed S A 2019 Effect of biodiesel fuels on diesel engine emissions *Egypt. J. Pet.*
[9] Goga G, Chauhan B S, Mahla S K and Cho H M 2019 Performance and emission characteristics of diesel engine fueled with rice bran biodiesel and n-butanol *Energy Reports***5** 78–83
[10] Selaimia R, Beghiel A and Oumeddour R 2015 The Synthesis of Biodiesel from Vegetable Oil *Procedia - Soc. Behav. Sci.***195** 1633–8
[11] Issariyakul T and Dalai A K 2014 Biodiesel from vegetable oils *Renew. Sustain. Energy Rev.***31** 446–71
[12] Huang D, Zhou H and Lin L 2012 Biodiesel: an Alternative to Conventional Fuel *Energy Procedia***16** 1874–85
[13] Roy M M, Wang W and Bujold J 2013 Biodiesel production and comparison of emissions of a DI diesel engine fueled by biodiesel–diesel and canola oil–diesel blends at high idling operations *Appl. Energy***106** 198–208
[14] Rajesh Kumar B, Saravanan S, Rana D and Nagendran A 2016 A comparative analysis on combustion and emissions of some next generation higher-alcohol/diesel blends in a direct-injection diesel engine *Energy Convers. Manag.***119** 246–56
[15] Atmanli A 2016 Comparative analyses of diesel-waste oil biodiesel and propanol, n-butanol or 1-pentanol blends in a diesel engine *Fuel***176** 209–15
[16] Redel-Macías M D, Pinzi S, Leiva-Candia D E, López I and Dorado M P 2017 Ternary blends of diesel fuel oxygenated with ethanol and castor oil for diesel engines *Energy Procedia***142** 855–60
[17] EL-Seesy A I, Hassan H and Kosaka H 2019 Improving the Performance of a Diesel Engine Operated with Jojoba Biodiesel-Diesel-n-Butanol Ternary Blends *Energy Procedia***156** 33–7
[18] Nanthagopal K, Ashok B, Saravanan B, Ramesh Pathy M, Sahil G, Ramesh A, Nurun Nabi M and Golam Rasul M 2019 Study on decanol and Calophyllum Inophyllum biodiesel as ternary blends in CI engine *Fuel***239** 862–73
[19] Babu D and Anand R 2017 Effect of biodiesel-diesel-n-pentanol and biodiesel-diesel-n-hexanol blends on diesel engine emission and combustion characteristics *Energy***133** 761–76
[20] Qi D H, Chen H, Geng L M, Bian Y Z and Ren X C 2010 Performance and combustion characteristics of biodiesel–diesel–methanol blend fuelled engine *Appl. Energy***87** 1679–86
[21] Gnanaprakasam A, Sivakumar V M, Surendhar A, Thirumarimurugan M and Kannadasan T 2013 Recent Strategy of Biodiesel Production from Waste Cooking Oil and Process Influencing Parameters: A Review *J. Energy***2013** 1–10
[22] Chai M, Tu Q, Lu M and Yang Y J 2014 Esterification pretreatment of free fatty acid in biodiesel production, from laboratory to industry *Fuel Process. Technol.***125** 106–13

Computational fluid dynamics (CFD) modelling of hybrid photovoltaic thermal system

Ashish Saurabh[1], Deepali Atheaya[2], Anil Kumar[3]
[1, 2]Department of Mechanical and Aerospace Engineering, Bennett University,
Tech Zone – II, Greater Noida, 201310, UP, India
[3]Department of Mechanical Engineering, Delhi Technological University, Delhi, India
[1]Corresponding author
E-mail: [1]*as6723@bennett.edu.in*, [2]*deepali.atheaya@bennett.edu.in*, [3]*anilkumar76@dtu.ac.in*

Received 14 October 2019; accepted 21 October 2019
DOI https://doi.org/10.21595/vp.2019.21098

Check for updates

Copyright © 2019 Ashish Saurabh, et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract. In this paper an attempt has been made to simulate and evaluate the distribution of temperature and heat flux for the hybrid photovoltaic thermal system with CFD (computational fluid dynamics) module in ANSYS 19.1 software. The simulation was carried to determine the temperature and heat flux across the different layers of HPVT (hybrid photovoltaic thermal) system. The temperature and heat flux were found to gradually decrease from the top glass layer to the bottom tedlar layer. The temperature varied from 36.4 °C (ambient) to 26.8 °C across the layers from top glass layer to bottom tedlar layer. The heat flux of (561.2 W/m^2) at the glass layer due to solar insolation was also found to gradually decrease with the increasing thickness. The reduction in temperature and heat flux with increasing thickness is attributed to fluid flowing beneath the system. The water flowing beneath the tedlar layer takes the excess heat out of the photovoltaic layer, thereby increasing the efficiency of the hybrid photovoltaic thermal system.

Keywords: computational fluid dynamics, hybrid photovoltaic thermal system, temperature distribution.

Nomenclature

T	Temperature (°C)
T_a	Ambient temperature (°C)
T_g	Glass temperature (°C)
T_i	Insulation temperature (°C)
T_c	Solar cell temperature (°C)
U_{ca}	Overall heat transfer coefficient from solar cell to ambient through glass cover (W/m^{2}k)
U_{cp}	Overall heat transfer coefficient from solar cell to plate (W/m^{2}k)
A	Area (m^2)
A_{rm}	Receiver area covered by PV module (m^2)
A_{pv}	Aperture area over PV module (m^2)
G_t	Total radiation (W/m^2)
A_{pf}	Area of absorber below PV module
ρ	Density (kg/m^3)
v	Velocity (m/s)
S_{mx}	Surface stress in x direction (N/m^2)
τ	Viscous stress (N/m^2)
u	Momentum in x direction (Nm)
CFD	Computational fluid dynamics
PVT	Photovoltaic thermal
HPVT	Hybrid photovoltaic thermal system
R	Thermal resistance

a	Air
g	Glass
s	Solar cell
t	Tedlar
ab	Absorber
w	Water
ins	Insulator
c	Conductive resistance
con	Convective resistance
r	Reflectivity
α	Absorptivity
τ	Transmittivity
β	Packing factor
F'	Flat plate collector efficiency factor

1. Introduction

Hybrid Photovoltaic Thermal System is the most promising emerging solution for current energy scenario. The efficiency of conventional solar PV decreases with increasing temperature. Solar cell efficiency decreases at the rate of 0.45 % per °C for silicon cells. Panel temperature can reach 70 °C in hot summer days. This means panels will produce up to 20 % less power compared to their rated power capacity at 25 °C. To overcome this, hybrid solar technology was developed. The hybrid system has the provision to capture the excess heat rejected by the solar PV and convert it into thermal energy. The efficiency of the hybrid system can reach up to 70 % [1, 2]. Khelifa et al. [3] performed a simulation analysis on a HPVT rectangular system with photovoltaic cells on the outside and air as heat transfer fluid. It was observed that the temperature of solar cells decreased due to convective heat transfer between the cells and the fluid medium and the fluid temperature varied from 20 °C to 40 °C between the inlet and outlet. Waeli et al. [4] modelled PVT system using nano fluid coolant and phase changing materials. The overall PVT efficiency was found to be 85.7 %. Hemmer et al. [5] numerically investigated heat transfer and flow dynamics in inclined PV panels at uniform heat flux of 510 W/m^2. The heat transfer was found low for large channel inclination angles. Edgar et al. [6] used a CFD finite element model to investigate the temperature of PV cell using basic material properties. Li and Jing [7] investigated the photovoltaic and thermal system performances of a 3D concentrated PVT (photovoltaic thermal) model using CFD and TRNSYS. The thermal energy efficiency gained its peak value against an optimal flow velocity different solar cell. The total energy efficiency was lower and exergy efficiency was higher for relatively low flow velocities. Zogou and Stapountzis [8] analyzed flow and heat transfer inside a PVT collector for building application. The flowrate and heat transfer characteristics of the backsheet proved to be most important factor for the performance of PV façade based on CFD simulations.

2. System description

The Fig. 1 shows the cross-sectional view of hybrid PVT. Solar insolation falls to the top glass layer. The heat flux is transmitted indirectly from the back side of solar cells. There is direct transmission of thermal energy from the non-packing area to the metal absorber plate by conduction. Further, there is transfer of thermal energy from absorber plate to fluid (water). The working fluid (water) is supplied from household to the hybrid collector where it gets heated from the thermal heat transferred by absorber plate through convection. The hybrid photovoltaic system helps reducing space, on site electricity generation and thermal heat generation. It also allows for the most efficient utilization of roof space. The various layers across the system are as follows:

1) Glass cover – this provides glazing and transmits radiation to the absorber. It also prevents radiative and convective loss from the surface. Solar insolation is incident on this layer. The thickness and other system properties are given in Table 1.

2) PV cell layer – below the glass cover lies the PV cells which convert solar energy to electricity by photovoltaic effect.

3) eva – eva is acronym used for ethyl vinyl acetate used to protect solar panel from extreme weather condition, moisture and dirt penetration.

4) Tedlar – tedlar is a protective sheet covering PV cells.

5) Absorber – this is composed of copper pipes carrying water to take the extra heat from the PV layer.

6) Insulation – it is done to reduce the heat losses to the surrounding.

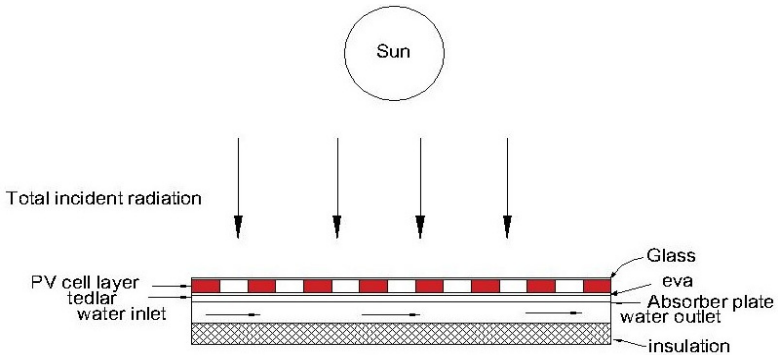


Fig. 1. Cross sectional view of Hybrid PVT

3. Energy analysis using governing equations

Following assumptions are made for writing the energy balance equations for the hybrid PVT collector system:

1) The system is in quasi steady state.

2) The ohmic losses in solar cells are negligible.

3) There is no temperature variation across thickness of solar cells and glass material.

4) Heat capacity of solar cell and glass cell is negligible.

5) One dimensional heat flow has been considered.

6) Transmittivity of eva is almost 100 %.

CFD i.e. computational fluid dynamics uses the following laws of conservation of mass, momentum and energy [9].

Continuity equation:

$$\frac{\partial \rho}{\partial t} + \vec{\nabla} \cdot (p\vec{v}) = 0. \tag{1}$$

Momentum conservation equation:

$$\rho \frac{\partial u}{\partial t} = \frac{\partial(-p + \tau_{xx})}{\partial x} \frac{+\partial \tau_{yx}}{\partial y} + \frac{\partial \tau_{zx}}{\partial z} + s_{Mx}. \tag{2}$$

Navier Stokes equation:

$$\frac{\partial \vec{v}}{\partial t} + (\vec{v} \cdot \vec{\nabla})\vec{v} = -\frac{\nabla p}{\rho} + \frac{u}{\rho}(\vec{\nabla} \cdot \vec{\nabla})\vec{v}. \tag{3}$$

Fig. 1 shows the cross-sectional view of hybrid PVT collector and Fig. 2 shows the corresponding thermal circuit with various resistances across the layers. The energy balance

equation for each layer of hybrid PVT collector (Atheaya et al. [7]) is as follows:

– For the PV module:

$$r\alpha_c\tau_g\beta_c G_t A_{pv} = \left[U_{ca}(T_c - T_a) + U_{cp}\left(T_c - T_p\right)\right]A_{PV} + r\eta G_t A_{PV}, \tag{4}$$

where $r\alpha_c\tau_g\beta_c G_t A_{pv}$ – rate of solar radiation available on PV cell layer, $U_{ca}(T_c - T_a)A_{PV}$ – rate of thermal energy loss from PV cell layer to ambient through glass, $U_{cp}\left(T_c - T_p\right)A_{PV}$ – rate of thermal energy loss from PV cell layer to absorber plate, $r\eta G_t A_{PV}$ – rate of generated electrical energy from PV cell layer.

– For the fluid flowing below the absorber:

$$\dot{m}_f c_f\left(T_{fo} - T_{fin}\right) = F' h_{pf}\left(T_f - T_p\right)A_{pv}, \tag{5}$$

where $\dot{m}_f c_f\left(T_{fo} - T_{fin}\right)$ – rate of energy carried away by flowing fluid, $F' h_{pf}\left(T_f - T_p\right)A_{pv}$ – rate of available energy to fluid.

– For absorber tube:

$$r\alpha_p\tau_g^2(1-\beta_c)G_t A_{pv} + U_{cp}\left(T_c - T_p\right)A_{PV} = F' h_{pf}\left(T_p - T_f\right)A_{PV}, \tag{6}$$

where $r\alpha_p\tau_g^2(1-\beta_c)G_t A_{pv}$ – rate of solar radiation available on absorber plate through non packing area of PV module, $U_{cp}\left(T_c - T_p\right)A_{PV}$ – rate of thermal energy transferred from solar cell to absorber plate, $F' h_{pf}\left(T_p - T_f\right)A_{PV}$ – rate of available energy to the fluid flowing below the absorber.

– The total resistance across the hybrid PVT system as per Fig. 2 will be given by:

$$R_{tot} = R_{a,c} + R_{g,c} + R_{s,c} + R_{t,c} + R_{ab,c} + R_{w,con} + R_{in,c},$$

$$R_{tot} = \frac{1}{h_1 A_g} + \frac{L_1}{k_1 A_g} + \frac{L_2}{k_2 A_s} + \frac{L_3}{k_3 A_t} + \frac{L_4}{k_4 A_{ab}} + \frac{1}{h_w A_{ab}} + \frac{L_5}{k_5 A_{in}}, \tag{7}$$

$$\dot{Q} = \frac{T_{a1} - T_{a2}}{R_{totW}}. \tag{8}$$

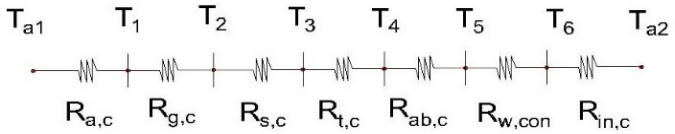


Fig. 2. Thermal resistance circuit diagram

4. Methodology

The methodology followed is shown schematically in Fig. 3. The geometric model for the hybrid system is generated using ANSYS software design modeller. The mesh is carried using ANSYS mesh module. The heat transfer phenomenon between different layers of hybrid photovoltaic module is modelled using the system coupling of Ansys transient thermal and Ansys FLUENT software. The fluid solid interface was created in FLUENT and simulated in transient thermal module. The thermal modelling of hybrid PVT system is analyzed with energy balance using the equations 1-8 written above. The temperature and heat flux variation of different layers of the hybrid system can be predicted.

5. Results and discussion

The design and operating parameters for the hybrid photovoltaic thermal system has been shown in Table 1. The meshing of the system was done for optimum value of skewness and aspect

ratio. The momentum, continuity and energy equation was solved using Fluent. Simple pressure velocity coupling with laminar flow model was chosen [10].

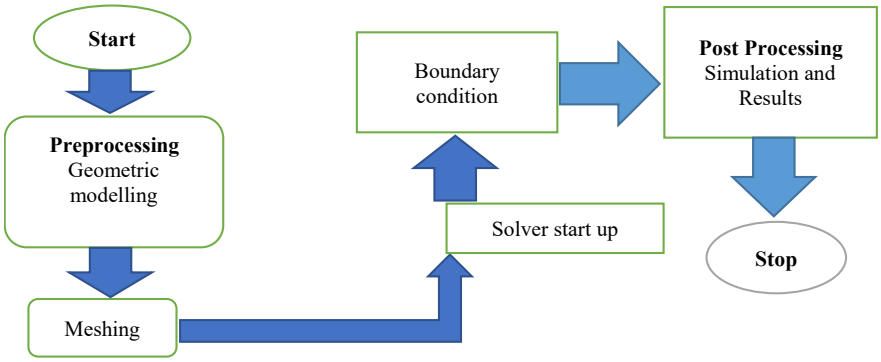


Fig. 3. Flowchart showing CFD simulation

Table 1. Design and operation parameters of hybrid PVT

Hybrid solar parameters	Values
Module type	monocrystalline
Area of PV module	1×1 m
Thickness of glass cover, solar cell, tedlar and back insulation	0.003 m, 0.0003 m, 0.0005 m, 0.05 m
Conductivity of glass cover, solar cell, tedlar and back insulation	0.98 W/m°C, 148 W/mK, 0.333 W/mK, 0.035 w/mK
Transmittivity of glass	0.95
Absorptivity of tedlar and solar cell	0.5, 0.9
Packing factor of solar cell	0.9
Ambient temperature and solar intensity	37 °C, 800 W/m^2

The governing equations for the CFD calculations are energy equations based on laminar model for the fluid flow through the domain. The model was first created in transient thermal and temperature variation was calculated across different layers as shown below in the Fig. 4. For CFD simulation, pressure-based solver was activated. The heat flux of 800 W/m^2 on the top layer due to solar radiation was simulated across the hybrid PVT system.

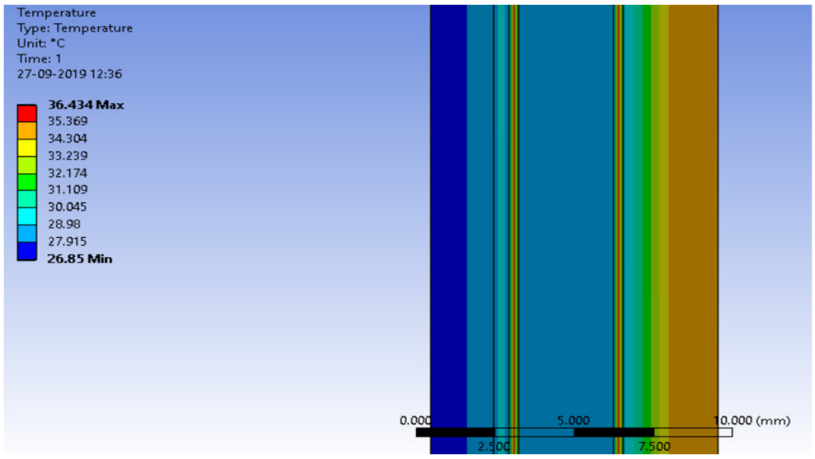


Fig. 4. Temperature variation across various layers of hybrid PVT

The fluid flow and heat transfer characteristics were studied with the ANSYS 19.1 Fluent through computational fluid dynamics calculations. It was observed that the temperature varied from 36.4 °C (ambient) to 26.8 °C across the layers from top to bottom as shown in Fig. 4. The decrease in temperature is due to heat taken away by water flowing beneath the tedlar layer and conductive heat flow across the various layers. Similarly, the variation of heat flux across the

hybrid photovoltaic thermal system is shown in Fig. 5. The heat flux was found to gradually decrease with the increasing thickness across the layers, reaching the minimum value at the bottom layer (tedlar). The gradual reduction in heat flux distribution is attributed to conductive resistances of different layers and convective resistance of fluid.

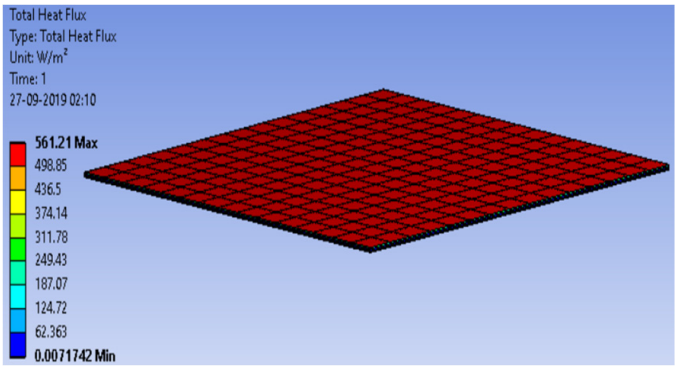


Fig. 5. Heat flux distribution across the hybrid PVT

6. Conclusions

Based on present studies following conclusions have been made.

1) The temperature varied from 36.4 °C (ambient) to 26.8 °C across the layers from top to bottom. The decrease in temperature is due to heat taken away by water flowing beneath the tedlar layer and conductive heat flow across the various layers. This will lead to increase in efficiency of the hybrid PVT system.

2) The heat flux (800 W/m^2) at the glass layer by solar insolation is gradually decreasing with the increasing thickness of the layers. reaching the minimum value at the bottom layer.

3) The efficiency of the hybrid PVT system will increase due to decrease in temperature across PV cell layer. Also, the heat taken by water flowing beneath the tedlar can be used for heating purpose.

References

[1] **Brooks Adria E.** Solar energy: photovoltaics. future energy (second edition). Improved, Sustainable and Clean Options for our Planet, Vol. 4, Issue 3, 2014, p. 383-404.

[2] **Abdulyan M., Chiasson A.** Thermal study of hybrid photovoltaic-thermal (PVT). Energy Procedia, Vol. 141, Issue 3, 2017, p. 102-108.

[3] **Khelifa A., et al.** Analysis of a hybrid solar collector photovoltaic thermal (PVT). Energy Procedia, Vol. 74, 2015, p. 835-843.

[4] **Al-Waeli A. H. A., et al.** Modelling and experimental validation of a PVT system using nanofluid coolant and nano-PCM. Solar Energy, Vol. 177, Issue 1, 2019, p. 178-191.

[5] **Hemmer C., et al.** Early development of unsteady convective laminar flow in an inclined channel using CFD: Application to PV panels. Solar Energy, Vol. 146, 2017, p. 221-229.

[6] **Edgar R., et al.** A computational fluid dynamic study of PV cell temperatures in novel platform and standard arrangements. Solar Energy, Vol. 144, Issue 1, 2017, p. 203-214.

[7] **Li Yang, Jing Dengwei** Investigation of the performance of photovoltaic/thermal system by a coupled TRNSYS and CFD simulation. Solar Energy, Vol. 143, 2017, p. 100-112.

[8] **Zogou Olympia, Stapountzis Herricos** Flow and heat transfer inside a PV/T collector for building application. Applied Energy, Vol. 91, Issue 1, 2012, p. 103-115.

[9] **Versteeg Malalasekra** An introduction to Computational Fluid Dynamics. Pearson, 2009.

[10] **Atheaya D., et al.** Analytical characteristic equation for partially covered photovoltaic thermal (PVT) compound parabolic concentrator (CPC). Solar Energy, Vol. 111, 2015, p. 176-185.

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/337722090

Degradation of dyes using biologically synthesized silver and copper nanoparticles

Article · December 2019

DOI: 10.1016/j.enmm.2019.100278

CITATIONS
0

READS
27

3 authors, including:



Arpita Roy
Delhi Technological University
62 PUBLICATIONS **88** CITATIONS
SEE PROFILE



Navneeta Bharadvaja
Delhi Technological University
48 PUBLICATIONS **148** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Green synthesis of nanoparticles View project

Molecular docking analysis of plumbagin View project

All content following this page was uploaded by Arpita Roy on 16 December 2019.

The user has requested enhancement of the downloaded file.

Environmental Nanotechnology, Monitoring & Management 13 (2020) 100278

Contents lists available at ScienceDirect

Environmental Nanotechnology, Monitoring & Management

journal homepage: www.elsevier.com/locate/enmm

Degradation of dyes using biologically synthesized silver and copper nanoparticles

Swati Raina, Arpita Roy, Navneeta Bharadvaja*

Plant Biotechnology Laboratory, Department of Biotechnology, Delhi Technological University, Delhi, India

ARTICLE INFO

Keywords:
Centella asiatica
Silver nanoparticles
Copper nanoparticles
Photocatalytic dye degradation
Catalytic dye degradation

ABSTRACT

Nanoparticle synthesis using plants have huge prospective as it is environment-friendly, cost-efficient, reproducible and energy efficient method as compared to physical or chemical synthesis. In the present study, green synthesis of silver and copper nanoparticles was done using *Centella asiatica*. Characterization of synthesized nanoparticles was done by using UV–vis spectroscopy and SEM. UV–vis gave characteristic SPR peak for silver and copper at 430 nm and 323 nm respectively. SEM images reveal spherical shape of synthesized silver and copper nanoparticles. Further comparative study of photocatalytic and catalytic degradation of methyl red, methyl orange, phenol red and eosin y using both the nanoparticles were done. CuNPs exhibited potential catalytic activity in methyl red (98.49 %), methyl orange (98.84 %) and phenyl red (99.62 %) respectively under optimal conditions.

1. Introduction

Nanotechnology is an emerging area and several technological advancements have been possible in the fields of medicine, catalysis and environmental clean-up, using nanoparticles. Nanoparticles possess remarkable properties due to their specific qualities like size, shape, and distribution as compared to bigger particles. Nanoparticles possess remarkable catalytic, optical, magnetic, electronic and thermal properties. Amongst various noble metals acting as nanoparticles most explored and potential nanoparticle that has been recognized for a long time is silver and copper nanoparticle. Silver and copper nanoparticles exhibit excellent catalytic properties. Silver is a common antimicrobial agent, owing to its strong toxicity against a various microorganism, silver-based compounds have been broadly used in many bactericidal applications (Ravindran et al., 2013). Silver nanoparticles has emerged itself as potential antifungal agent against *Candida albicans*, *Candida nonalbicans* and *Candida tropicalis* (Roy, 2017). Copper nanoparticles have gained attention due to lower cost, easily available and properties similar to other metallic nanoparticles (Nagar and Devra, 2018). Copper nanoparticles are extensively utilized in sensors, heat transfer fluids, antimicrobial agents and catalysis (Ahmed et al., 2016). A variety of methods are available for nanoparticles synthesis like chemical reduction, sol gel, etc. unfortunately these methods uses harmful chemicals or required high energy, which are rather difficult and generate harmful waste (Mittal et al., 2014). Biosynthetic methods employing green chemistry exploit use of either microorganisms or plant extracts which have emerged as easy and feasible alternative to other methods. Green synthesis offers advantage of being environment friendly, single step process, cost-effective, reproducible and often results in stable materials (Bogireddy et al., 2015). Plant extract contains various phytocompounds like proteins, flavonoids, polyphenols, terpenoids, etc., which act as reducing and capping agents (Ahmed et al., 2016). Microorganisms can also be used for synthesis of nanoparticles but rate of synthesis is time-consuming as compared to plant drive synthesis (Mittal et al., 2014).

Organic dyes are generally used in industries like plastic, paper, pharmaceutical, cosmetic, food, and textile. They pose potential hazard to our environment and water bodies. Most commonly used organic dyes in industries includes methylene blue, methyl red, congo red, methyl orange, eosin Y, bromophenol blue, acridine orange, phenol red, etc. (Khan et al., 2016; Roy and Bharadvaja, 2019). Complete eradication of dyes before expulsion of industrial waste is necessary to minimize water pollution as dyes are highly stable in water owing to their complicated structure. Green approach utilizing suitable nanocatalyst in presence of sunlight or chemicals offers a potential method for elimination of dyes from environment.

In this study silver and copper nanoparticles were synthesized from *Centella asiatica* and their potential in catalytic and photo-catalytic dye degradation was studied. A rapid, eco-friendly and convenient method for copper and silver nanoparticle synthesis using leaf extract of *C.*

* Corresponding author.
E-mail address: navneetab@dce.ac.in (N. Bharadvaja).

https://doi.org/10.1016/j.enmm.2019.100278
Received 8 September 2019; Received in revised form 5 November 2019; Accepted 2 December 2019
2215-1532/ © 2019 Elsevier B.V. All rights reserved.

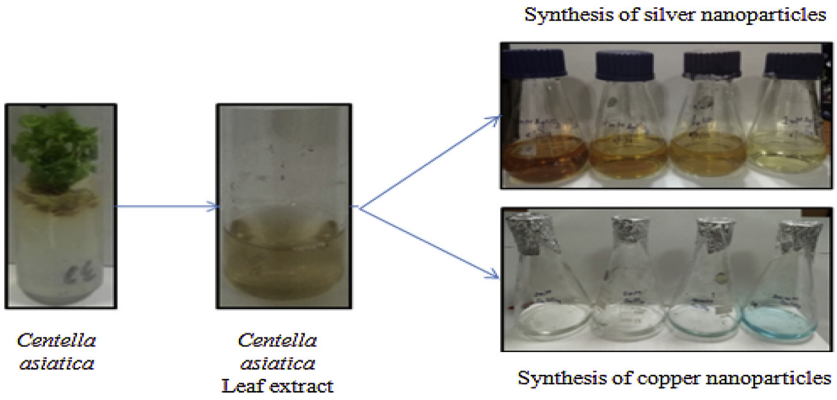


Fig. 1. Synthesis of silver and copper nanoparticles from *Centella asiatica.*

asiatica has been reported. Four different concentrations of silver nitrate and cupric sulphate were used for nanoparticles synthesis which were further utilized for catalytic and photocatalytic degradation of dyes.

2. Materials and methods

Centella asiatica was maintained in MS +1 mg/l BAP (Roy et al., 2016; Roy and Bharadvaja, 2017a, 2017b, 2017c). Fresh green leaves of *C. asiatica* were collected, weighed and boiled in distilled water (1gm/10 ml) in a sterilized 250 ml conical flask for 15 min at 60 °C. Extract was filtered using Whatman Filter Paper No. 1 to remove impurities and stored at 4 °C.

Four concentration of silver nitrate (1 mM, 2 mM, 3 mM and 4 mM) and cupric sulphate (1 mM, 5 mM, 20 mM and 100 mM) were used for nanoparticles synthesis. Silver nanoparticles synthesis was indicated by yellowish-brown color formation and copper nanoparticles synthesis was indicated by change in color from colorless to bluish green after addition of plant extract (Fig. 1). Further confirmation of synthesized nanoparticles was done by UV–vis s spectrophotometer (PerkinElmer UV/VIS Lambda 365). Distilled water was taken as blank for both the nanoparticles. Silver nanoparticles were characterized at a wavelength range of 300 – 700 nm and copper nanoparticles at 200 – 700 nm. SEM was used for the shape analysis of synthesized nanoparticles.

Photocatalysis: 1 ml of 1 mM Methyl orange (MO), Methyl red (MR), Phenol red (PR) and Eosin Y(EY) was added to 1 ml of silver nanoparticle (1 mg/ml) synthesized from different concentration of silver nitrate (1 mM, 2 mM, 3 mM and 4 mM) solution. Similarly, 1 ml of 1 mM Methyl orange (MO), Methyl red (MR), Phenol red (PR) and Eosin Y(EY) was added to 1 ml of copper nanoparticles(1 mg/ml) synthesized from different concentration of cupric sulphate (1 mM, 2 mM, 3 mM and 4 mM). In both cases volume was made up to 5 ml by adding distilled water to the mixture. Solution was kept for stirring in dark for 2 h to achieve equilibrium between dye and nanoparticles. After dark incubation, solution was kept in sunlight for 0 – 24 h to check degradation rate. Dye degradation process was analyzed by UV–vis spectrophotometer. Solution of dye + water was taken as blank for both nanoparticles. Experiments were repeated thrice and mean percentage value was recorded.

Chemical catalysis: 1 ml of 100 mM sodium borohydride solution was added to 1 ml of 1 mM methyl orange (MO), Methyl red (MR), Phenol red (PR) and Eosin Y (EY) respectively. 1 ml of silver nanoparticles (1 mg/ml) synthesized from different concentration of silver nitrate was added to each dye with sodium borohydride solution and stirred. Similarly, 1 ml of copper nanoparticles synthesized (1 mg/ml) from different concentration of cupric sulphate was added to each dye with sodium borohydride solution and stirred. Volume was made up to 5 ml using distilled water. Degradation was indicated by decolorization of solution and analyzed using UV–vis absorption spectrophotometer. Rate of dye degradation was recorded up to 3 h, after which no significant reduction was observed. Dye + water solution was taken as blank for both nanoparticles. Experiments were repeated thrice and mean percentage value was recorded. Instrument was calibrated with optimum wavelength. Limit of detection and limit of quantification for silver nanoparticle was 0.5 mM and 1.6 mM and for copper nanoparticles it was 0.6 mM and 1.7 mM respectively.

3. Results and discussion

Silver nanoparticles were synthesized from *Centella asiatica* leaf extract by adding different concentrations of $AgNO_3$ (1 mM, 2 mM, 3 mM, 4 mM). Colorless to yellowish brown color indicates synthesis of silver nanoparticles. By increasing silver nitrate concentration, intensity of yellowish-brown color increased which indicates high amount of nanoparticle formation. Further confirmation of nanoparticles synthesis was done using UV–vis spectrophotometer. UV–vis absorption spectra of samples showed highlighted peak in the range of 330 – 550 nm which is characteristic of silver nanoparticles (Ramesh et al., 2015; Roy and Bharadvaja, 2017a, 2017b, 2017c; Roy and Bharadvaja, 2019). Due to SPR phenomenon, AgNPs exhibited strong absorption peak in the visible range and maximum SPR peak was observed at 430 nm (Fig. 2).

Copper nanoparticles were synthesized using different concentrations of cupric sulphate (1 mM, 5 mM, 20 mM and 100 mM). Immediate color change was observed from colorless to bluish green, which indicates formation of copper nanoparticles. Change from light blue to bluish green from 1 mM to 100 mM cupric sulphate indicates high amount of nanoparticle formation at 100 mM. Absorbance peak of copper nanoparticles was in the range of 250 – 360 nm. Absorbance of nanoparticle ranged from 0.8 to 3.05 for 1 mM, 5 mM, 20 mM cupric sulphate whereas in case of 100 mM cupric sulphate absorption ranged from 0.9 to 1.8 because of SPR effect, more the concentration of nanoparticles lesser the size. Maximum SPR peak for copper nanoparticles was observed at 323 nm. Karthik et al. (2017) reported SPR peak of silver nanoparticles at 405 – 434 nm at different concentration of plant extracts whereas Roy and Bharadvaja (2017a, 2017b, 2017c) obtained SPR peak of silver nanoparticles at 446 nm. Nasrollahzadeh and Sajadi (2015) found SPR peak at 350 nm for copper nanoparticles.

SEM images of silver and copper nanoparticles reveal spherical shape and agglomerated nanoparticles with estimated size range of 30 – 50 nm and 20 – 30 nm respectively (Fig. 3). Synthesized silver nanoparticle size was comparable with previous studies on *Aloe vera* (Tippayawat et al., 2016) and Crocus sativus (Bagherzade et al., 2017). Nayak et al. (2016) reported silver nanoparticle synthesis from *F. benghalensis* and *A. indica* of size 85.95 nm and 90.13 nm respectively. Ramesh et al. (2015) reported spherical shape of silver nanoparticles synthesized from *Emblica officinalis* ranging from 10 to 70 nm. Roy and Bharadvaja (2019) reported spherical silver nanoparticles of size 55 nm. Size of synthesized copper nanoparticles was 28.73 nm as reported by Khan et al. (2016). Chung et al. (2017) reported copper

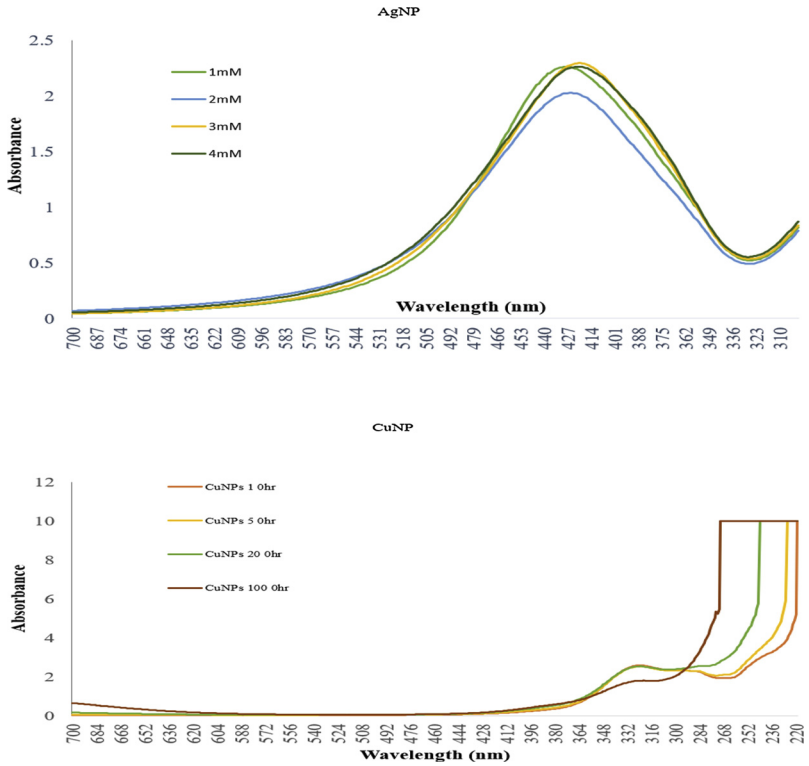


Fig. 2. UV–vis spectra of AgNPs and CuNPs.

nanoparticles of size 31 nm from Eclipta prostrata.

Synthesized silver nanoparticles were examined for dye degradation potential. Absorption spectrum at different conditions was recorded. For photocatalytic degradation of dyes solar irradiation was used. Degradation was in order of Eosin Y > Phenol Red > Methyl Orange > Methyl Red. Maximum degradation was observed in phenol red and eosin Y and minimum in methyl orange. Silver nanoparticles synthesized from 1 mM concentration showed highest degradation of 93.88 % in case of Eosin Y (Fig. 4). For catalytic dye degradation $NaBH_4$ was used along with silver nanoparticles. Solution of methyl orange, methyl red and phenol red turned colorless in presence of $NaBH_4$ which indicate reduction of dyes. Silver nanoparticles synthesized from 1 mM concentration showed highest degradation of 97.88 % in phenol red. Methyl orange was degraded upto 84.38 % at 1 mM followed by 3 mM and 2 mM i.e. 77.69 % and 74.25 % respectively and least degradation was recorded at 4 mM concentration. Maximum degradation of methyl red was observed at a concentration of 4 mM and 1 mM i.e. 82.9 % and 76.2 % respectively. Eosin Y was least degraded at all concentrations and 14 % degradation was recorded at 1 mM concentration. 1 mM and 3 mM concentrations showed best results of dye degradation up to 3 h. No further degradation was observed after 3 h. MO, MR, PR became colorless after 3 h at all concentrations. Percentage of dye degradation at 3 h was PR > MO > MR > EY (Fig. 5). Catalytic degradation process showed better degradation rates in less time as compared to the photcatalytic degradation. This could be due to the ability of $NaBH_4$ to accelerate the degradation process in less time as compared to photocatalytic method where $NaBH_4$was absent.

Copper nanoparticles showed potential degradation by photocatalytic and catalytic method. Dye degradation percentage was determined by UV–vis spectra at different time intervals. Maximum photocatalytic degradation of methyl orange and phenol red at concentration of 100 mM was 73.00 % and 89.04 % respectively. Maximum catalytic degradation of methyl orange was obtained at concentration of 100 mM i.e. 99.84 % (Figs. 6 and 7). Eosin Y showed 12.18 % degradation at 100 Mm concentration and dye precipitated at the bottom of test tube observed 5 − 7 min after addition of nanoparticles).

Silver and copper nanoparticles showed potential results of dye degradation with sodium borohydride and moderate results with photocatalytic method. Joseph and Mathew (2015) reported methyl orange and rhodamine B degradation using silver nanoparticles with sodium borohydride. Fatimah (2016) reported catalytic degradation of methyl orange using silver nanoparticles. Devi and Singh (2014) reported synthesis of CuO nanoparticles from *C. asiatica* and degraded methyl orange using photocatalytic method. They reported degradation of dyes

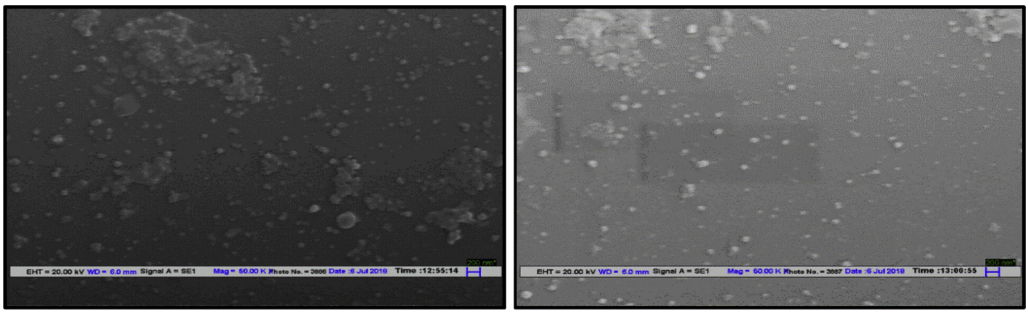


Fig. 3. SEM image of silver and copper nanoparticles.

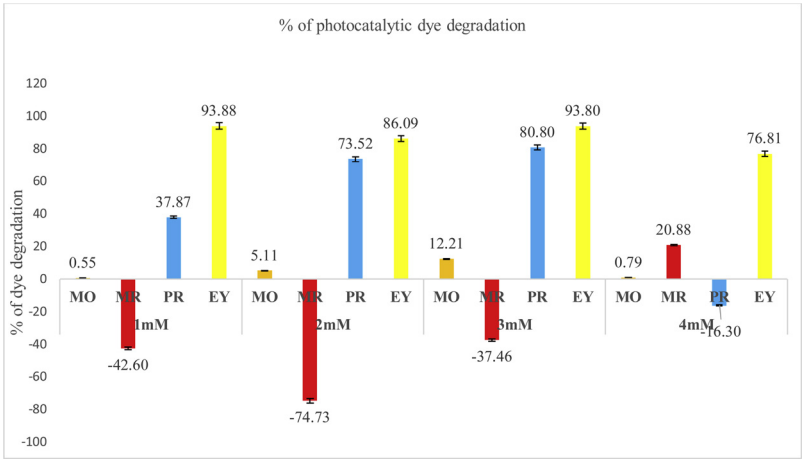


Fig. 4. Percentage of dye degradation by AgNps with photocatalytic method.

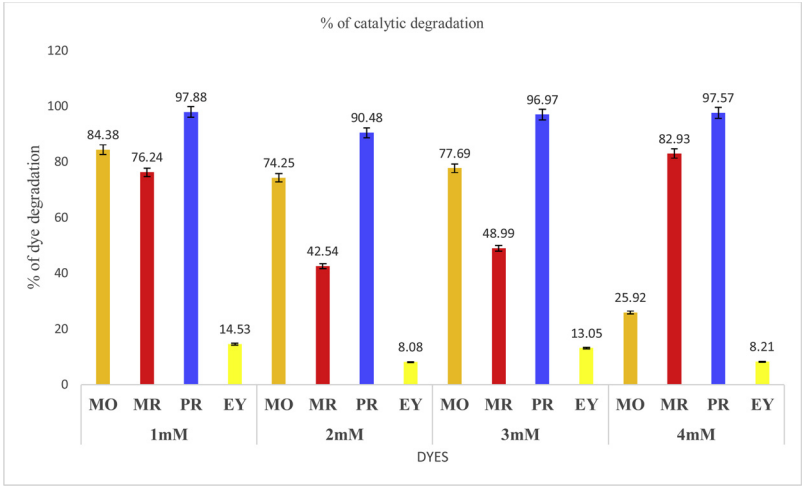


Fig. 5. Percentage of dye degradation by AgNps with catalytic method using NaBH4 over 3 h.

within 10 min of reaction and showed 66.66 % degradation of dye. In the present study 77.69 % dye degradation was achieved at 3 mM concentration of $AgNO_3$. Catalytic process utilizing copper nanoparticles degraded methyl red, methyl orange and phenol red upto 97 % in 5 min. MeenaKumari and Philip (2015) reported degradation of textile dyes like methylene blue, methyl orange and eosin Y using catalytic method within 20 min.

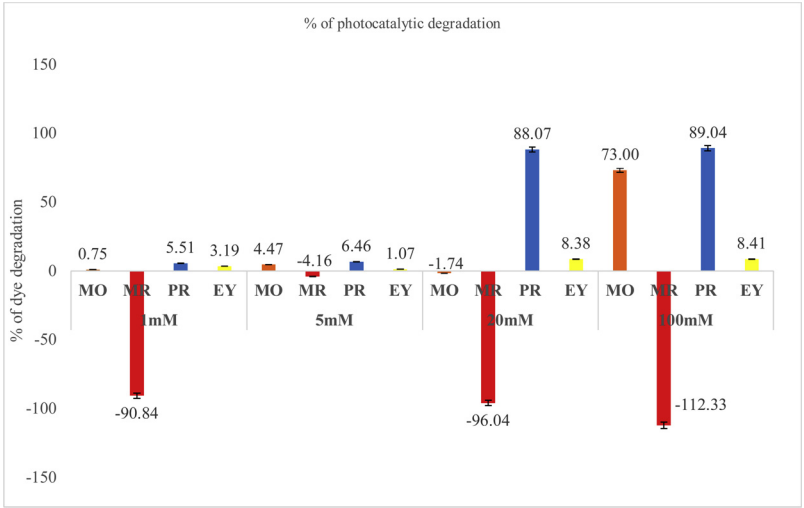


Fig. 6. Percentage of dye degradation by CuNps with photocatalytic method.

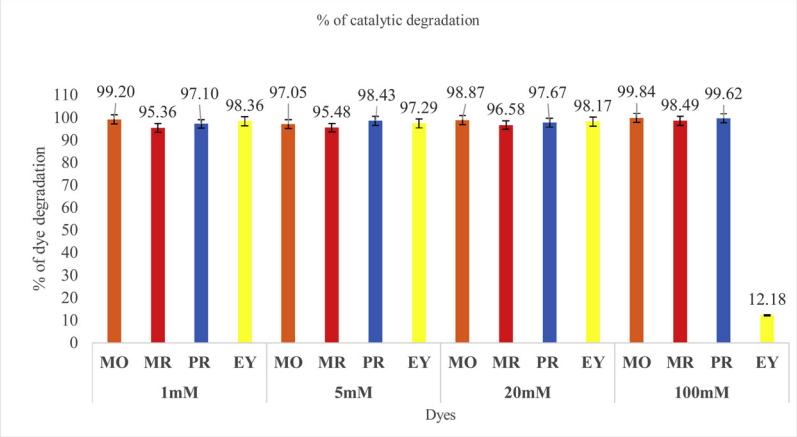


Fig. 7. Percentage of dye degradation by CuNPs with catalytic method using $NaBH_4$ over 15 min.

4. Conclusion

Release of organic dyes is a major concern for environment as these dyes reach municipal wastewater. Due to their high stability in water they are not degraded and end up in treated water which is used in agricultural fields. Textile dyes released in water bodies from any industry are posing threat to environment and mankind. Nanoparticles have proven vivid catalytic properties thus can be utilized for dye degradation. Metallic nanoparticles synthesized using plants have several advantages compared to other methods. In this study two different types of nanoparticles i.e. silver and copper were synthesized using four different concentrations of silver nitrate and cupric sulphate with the help of a medicinal plant *Centella asiatica*. Further these nanoparticles that are synthesized at different concentrations were used for the degradation of four different dyes. Degradation studies showed that among silver and copper nanoparticles, copper nanoparticles showed better results in degrading all the four dyes in a catalytic process. Copper nanoparticles are economical as compared to silver nanoparticles, therefore these can be effectively utilized for large scale degradation of these dyes.

Acknowledgement

We would like to acknowledge Delhi Technological University for providing all necessary facilities.

References

Ahmed, S., Saifullah, Ahmad, M., Swami, B.L., Ikram, S., 2016. Green synthesis of silver nanoparticles using Azadirachta indica aqueous leaf extract. J. Radiat. Res. Appl. Sci. 9 (1), 1–7.

Bogireddy, N.K.R., Anand, K.K.H., Mandal, B.K., 2015. Gold nanoparticles—synthesis by *Sterculia acuminata* extract and its catalytic efficiency in alleviating different organic dyes. J. Mol. Liq. 211, 868–875.

Bagherzade, G., Tavakoli, M.M., Namaei, M.H., 2017. Green synthesis of silver nanoparticles using aqueous extract of saffron (Crocus sativus L.) wastages and its antibacterial activity against six bacteria. Asian Pac. J. Trop. Biomed. 7 (3), 227–233.

Chung, I.M., Abdul Rahuman, A., Marimuthu, S., Kirthi, A.V., Anbarasan, K., Padmini, P., Rajakumar, G., 2017. Green synthesis of copper nanoparticles using *Eclipta prostrata* leaves extract and their antioxidant and cytotoxic activities. Exp. Ther. Med. 14 (1), 18–24.

Devi, H.S., Singh, T.D., 2014. Synthesis of copper oxide nanoparticles by a novel method and its application in the degradation of methyl orange. Adv Electron Electr Eng 4 (1), 83–88.

Fatimah, I., 2016. Green synthesis of silver nanoparticles using extract of *Parkia speciosa* Hassk pods assisted by microwave irradiation. J. Adv. Res. 7 (6), 961–969.

Joseph, S., Mathew, B., 2015. Microwave-assisted green synthesis of silver nanoparticles and the study on catalytic activity in the degradation of dyes. J. Mol. Liq. 204, 184–191.

Karthik, R., Govindasamy, M., Chen, S.M., Cheng, Y.H., Muthukrishnan, P., Padmavathy, S., Elangovan, A., 2017. Biosynthesis of silver nanoparticles by using *Camellia japonica* leaf extract for the electrocatalytic reduction of nitrobenzene and photocatalytic degradation of Eosin-Y. J. Photochem. Photobiol. B 170, 164–172.

Khan, H., Khalil, A.K., Khan, A., Saeed, K., Ali, N., 2016. Photocatalytic degradation of bromophenol blue in aqueous medium using chitosan conjugated magnetic nanoparticles. Korean J. Chem. Eng. 33 (10), 2802–2807.

MeenaKumari, M., Philip, D., 2015. Degradation of environment pollutant dyes using phytosynthesized metal nanocatalysts. Spectrochim. Acta A. Mol. Biomol. Spectrosc. 25 (135), 632–638.

Mittal, J., Batra, A., Singh, A., Sharma, M.M., 2014. Phytofabrication of nanoparticles through plant as nanofactories. Adv. Nat. Sci. Nanosci. Nanotechnol. 5 (4), 043002.

Nagar, N., Devra, V., 2018. Green synthesis and characterization of copper nanoparticles using Azadirachta indica leaves. Mater. Chem. Phys. 213, 44–51.

Nasrollahzadeh, M., Sajadi, S.M., 2015. Green synthesis of copper nanoparticles using *Ginkgo biloba* L. leaf extract and their catalytic activity for the Huisgen [3+ 2] cycloaddition of azides and alkynes at room temperature. J. Colloid Interface Sci. 457, 141–147.

Nayak, D., Ashe, S., Rauta, P.R., Kumari, M., Nayak, B., 2016. Bark extract mediated green synthesis of silver nanoparticles: evaluation of antimicrobial activity and antiproliferative response against osteosarcoma. Mater. Sci. Eng. C 58, 44–52.

Ramesh, P.S., Kokila, T., Geetha, D., 2015. Plant mediated green synthesis and antibacterial activity of silver nanoparticles using *Emblica officinalis* fruit extract. Spectrochim. Acta A. Mol. Biomol. Spectrosc. 142, 339–343.

Ravindran, A., Chandran, P., Khan, S.S., 2013. Biofunctionalized silver nanoparticles: advances and prospects. Colloids Surf. B Biointerfaces 105, 342–352.

Roy, A., 2017. Synthesis of silver nanoparticles from medicinal plants and its biological application: a review. Res. Rev. Biosci. 12 (4), 138.

Roy, A., Bharadvaja, N., 2019. Silver Nanoparticles Synthesis from *Plumbago zeylanica* and its dye degradation activity. Bioinspir. Biomim. Nan. https://doi.org/10.1680/jbibn.18.00036.

Roy, A., Bharadvaja, N., 2017a. Silver nanoparticles synthesis from a pharmaceutically important medicinal plant *Plumbago zeylanica*. MOJ Bioequiv Availab. 3 (5), 00046.

Roy, A., Kundu, K., Saxena, G., Kumar, L., Bharadvaja, N., 2016. Effect of different media and growth hormones on shoot multiplication of in vitro grown *Centella asiatica* accessions. Adv. Tech. Biol. Med. 4, 172.

Roy, A., Bharadvaja, N., 2017b. Qualitative analysis of phytocompounds and synthesis of silver nanoparticles from *Centella asiatica*. Innov. Tech. Agric. 1 (2), 88–95.

Roy, A., Bharadvaja, N., 2017c. Effect of different culture medias on shoot multiplication and stigmasterol content in accessions of centella asiatica. Int. J. Ayurved. Herbal Med. 7, 2643–2650.

Tippayawat, P., Phromviyo, N., Boueroy, P., Chompoosor, A., 2016. Green synthesis of silver nanoparticles in aloe vera plant extract prepared by a hydrothermal method and their synergistic antibacterial activity. PeerJ 4, e2589. https://doi.org/10.7717/peerj.2589.

View publication stats

Development of Reliable and High Responsivity ZnO-Based UV-C Photodetector

Jitesh Agrawal, Tejendra Dixit, I. A. Palani, and Vipul Singh

***Abstract*— This work is focused on the development of reliable and high-performance ZnO based deep UV (UV-C) photodetectors. Herein, hydrothermally synthesized ZnO honeycomb nanostructures were utilized to develop a high sensitivity UV-C photodetector. A systematic analysis of the device performance and stability in deep UV (250 nm) radiations has been performed. The as-synthesized device has shown high sensitivity of 2.4 × 10^5 and responsivity of 597 A/W, at 20 V applied bias. Further, we have demonstrated that the coating of Pt nanoparticles over ZnO nanostructures could significantly improve not only the device stability but also the response speed and device endurance in deep UV radiations. The proposed device is a promising contender for future deep UV detector based optoelectronic products.**

***Index Terms*— Deep UV photodetector, honeycomb, hydrothermal growth, nanostructures, ZnO.**

I. Introduction

IN THE recent years, a significant amount of efforts have been made towards the development of deep UV photodetectors due to their numerous applications such as flame detection, inter-satellite communication, missile launching systems and harmful UV detection [1]. Variety of wide band semiconducting materials viz. graphene QDs (3.8 eV), β-gallium oxide (4.9 eV), boron nitride (4.0-5.5eV) and zinc oxide have been explored for development of UV sensors etc. [2]–[8]. Among these materials, ZnO has been recognized as a potential metal oxide semiconducting material for the fabrication of UV photodetectors owing to its wide and direct bandgap of 3.37 eV and large exciton binding energy of 60 meV [9]. Further, ZnO is low cost, bio-compatible material with highly tunable structural and optical properties that has made the device suitable for large range of applications. However, the pristine ZnO is not sensitive to deep UV radiations. In order to make ZnO sensitive enough in this region various metals like Mg, Cd, Ge, Ga and Be have been doped [10]–[14]. The doping process itself is quite a complex process and increase the cost of device fabrication. Moreover, such devices have not shown noteworthy improvement in the device performance in the deep UV region. Another major issue with ZnO based deep UV sensors is their stability in deep UV radiations as when the devices are exposed to such high energy radiations there are high chances of material to degrade. Therefore, before commercializing such devices, it is highly required to examine the device reliability and stability in such harsh environmental conditions. However, as per our knowledge, the issue has been rarely discussed. The metallic nanoparticles which are comprehensively explored in the last decade and have demonstrated significant improvement in UV emission and sensing response of ZnO, could pave the way to improve the device stability and reliability in deep UV radiations.

In the present work we have used a simple and low-cost hydrothermal technique for the growth of ZnO honeycomb nanostructures by incorporating $Na_3C_6H_5O_7$ i.e. trisodium citrate in the precursor solution. The obtained honeycomb nanostructures have provided not only a larger surface area but also facilitated efficient charge transportation between the electrodes which was resulted in improved device performance. Further, the incorporation of Zinc interstitial in the ZnO lattice was confirmed by Raman and absorption spectra which was responsible for deep UV sensing response of the device. Moreover, to examine the device reliability in UV radiations, the samples were exposed to UV light for prolong duration of time and were intensively characterized by PL and I-V characterization. Finally, we have demonstrated that the Pt NPs coating over ZnO nanostructures is an effective way to prevent material degradation and to improve the device performance and endurance in UV-C radiation.

II. Experimental Section

The samples were prepared over 1 cm × 1 cm borosilicate glass substrates. The substrates were ultrasonicated in Acetone, IPA and Deionized (DI) water for ten minutes each, respectively. Meanwhile, a solution of a 5M concentration of $Zn(CH_3CO_2)_2$, monoethanolamine in 2-Methoxyethanol was prepared. The solution was spin-coated over the substrate at 3000 rpm for 30 seconds, then annealed at 240 °C for 5 minutes. Thereafter, a precursor solution of 0.1 M concentration of $Zn(NO_3)_2.6H_2O$ and $C_6H_{12}N_4$ were prepared and samples were placed in the solution for 4 Hrs. at 110 °C. Further, $Na_3C_6H_5O_7$ was added in the precursor solution as a surfactant for obtaining different morphologies of ZnO. Notably, the samples grown at 0 mM, 0.5 mM, 5 mM and 10 mM concentration of $Na_3C_6H_5O_7$ were named as Pristine ZnO, 0.5CA, 5CA and 10CA, respectively. After the

Manuscript received August 1, 2019; revised November 13, 2019; accepted November 29, 2019. Date of publication December 4, 2019; date of current version December 19, 2019. This work was supported by University Grants Commission (UGC), India, under Grant 3526/(NET-DEC. 2014). *(Corresponding author: Vipul Singh.)*

J. Agrawal and V. Singh are with the Molecular and Nanoelectronics Research Group (MNRG), Department of Electrical Engineering, IIT Indore, Indore 453552, India (e-mail: vipul@iiti.ac.in).

T. Dixit is with the Department of Physics, IIT Madras, Chennai 600036, India, and also with the Nano Functional Materials Technology Centre, IIT Madras, Chennai 600036, India.

I. A. Palani is with the Mechatronics and Instrumentation Laboratory, Department of Mechanical Engineering, IIT Indore, Indore 453552, India.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JQE.2019.2957584

0018-9197 © 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

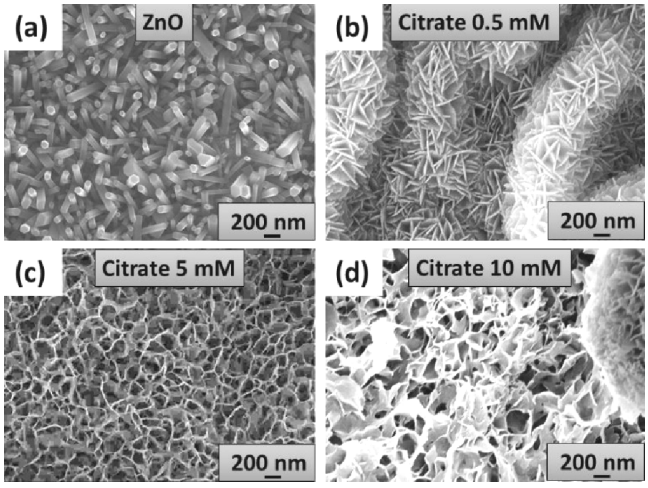


Fig. 1. FESEM images of (a) Pristine ZnO, (b) 0.5CA, (c) 5CA and (d) 10CA samples.

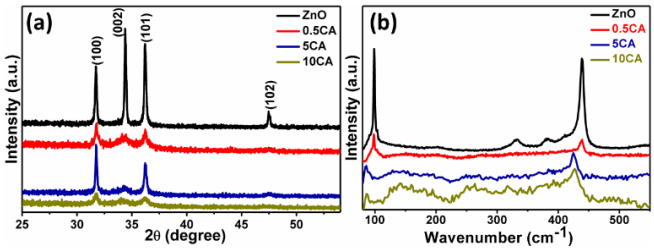


Fig. 2. (a) XRD plot and (b) Raman spectra of the pristine ZnO, 0.5CA, 5CA and 10CA samples.

hydrothermal growth, Ag electrodes of 50 μm channel length and 2 mm width were deposited using shadow masks in physical vapor deposition (PVD).

ZnO nanostructures (NSs) morphologies were characterized by Zeiss Supra-55, FESEM system. Rigaku Smartlab, Automated multipurpose XRD with Cu-Kα source has been used to investigate the samples' crystallinity. Raman analysis has been performed using Horiba Jobin Yuon HR 800 UV Raman Spectrometer at 488 nm wavelength. Photoemission spectra were obtained by Dongwoo Optron DM 500 i Photoluminescence (PL) analyzer. Cary 60 UV-Vis, Agilent Technologies was used to obtain absorption spectra of the samples. Keithley 2612A source meter was used for electrical characterizations. Incident light power was measured by TS2 LP 100, CNI laser, laser power meter.

III. Results and Discussions

The FESEM image in Fig. 1(a) has shown growth of ZnO nanorods when $Na_3C_6H_5O_7$ was not added in the precursor solution during hydrothermal growth. On addition of 0.5 mM and 5 mM concentration of $Na_3C_6H_5O_7$, growth of nanoplates and honeycomb structures takes place (shown in Fig. 1(b, c). However, as the $Na_3C_6H_5O_7$ level increased to 10 mM, ZnO precipitates were getting formed in the solution which were settled down over the glass substrate (fig. 1(d)).

The XRD plot in Fig. 2(a) has shown four sharp peaks at 31.73°, 34.40°, 36.22° and 47.47° corresponding

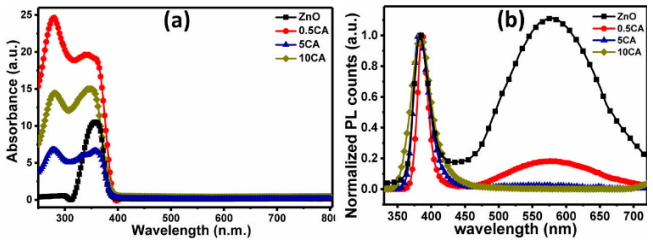


Fig. 3. (a) Absorption spectra and (b) PL spectra of the pristine ZnO, 0.5CA, 5CA and 10CA samples.

to (100), (002), (101) and (102) planes of Zinc oxide (as per JCPDS 01-089-7102). Notably, in the XRD plot of pristine ZnO, (002) diffraction peak has the highest intensity which indicated towards the preferential growth along the c-axis. However, XRD plot of 5CA has shown significant suppression in the (002) peak, suggesting that the growth was inhibited along c-axis. As a result, growth has taken place along nonpolar facets of ZnO to form nanoplates like structures. Moreover, the XRD plot of 10CA sample has shown significant suppression in all XRD peaks intensity, which could be assigned to lower crystallinity of ZnO nanostructures. The XRD results were well supported by FESEM images.

In Fig. 2(b) Raman spectra of pristine ZnO has shown two sharp peaks at 98.77 cm^{-1} and 438.52 cm^{-1} corresponds to E_2(low) and E_2(high) modes [15]. Further, with increasing $Na_3C_6H_5O_7$ concentration, the Raman peaks become broader and the intensity of the peaks were reduced, which indicates towards decreased crystallinity and/or increase in defect states. Moreover, Raman spectra of 5CA sample has revealed large shift in E_2(high) peak towards lower wavenumber (422.68 cm^{-1}), that could be assigned to the incorporation of Zinc interstitials in the ZnO lattice [16].

Moreover, with the purpose of studying absorption properties and to probe the presence of Zinc interstitials in the ZnO lattice, absorption spectroscopy analysis has been performed. The samples have not shown noteworthy absorption in visible and near infrared wavelength range. The absorption spectra of pristine ZnO shows single band edge peak centered at 357 nm. Interestingly, the other three samples have shown additional absorption peak around 270 nm. As the trisodium citrate is a strong reducing agent, it could reduce the Zinc ions in the growth solution, which could be incorporated in the ZnO lattice. These Zinc interstitials were ascribed to the absorption peak at 270 nm [17].

Further, photoluminescence (PL) analysis has been performed to investigate the luminescence properties and the defects present in the samples. For the PL analysis, samples were excited by 325 nm He-Cd laser source. All four samples have shown intense near band emission (NBE) peak, centered at 380 nm. Noteworthily, as the concentration of $Na_3C_6H_5O_7$ is increased in the precursor solution, the NBE peak has become broader, which could be ascribed to variation in local bandgap due to local strain in the nanostructures [18]. Additionally, a small tale of NBE peak, extending towards the visible region has been observed and was assigned to the Zinc interstitial defects [17]. Further, a separate broad peak, centered at 580 nm has been observed. This peak is

due to the mid-gap states formed due to the various oxygen vacancies related defects in the samples, known as deep-level emission (DLE) peak [19]. The PL spectra of pristine ZnO has shown intense DLE peak which signifies the presence of excessive oxygen vacancies related defect states. Interestingly, the intensity of DLE peak has been significantly suppressed in 0.5CA and almost vanishes in 5CA and 10CA samples which clearly indicated that defect density has been drastically reduced with the addition of trisodium citrate in the precursor solution used for hydrothermal growth of ZnO.

Further, to investigate the device photo-sensing response, I-V characterization has been performed and various device parameters like sensitivity, responsivity, specific detectivity, external quantum efficiency (EQE) and linear dynamic range (LDR) were calculated which can be explained as follows:

Photosensitivity (S): It is defined as the ratio of difference between photo current (I_{ph}) and dark current (I_d) to the dark current.

$$Sensitivity = \frac{I_{ph} - I_d}{I_d}$$

Photoresponsivity (R): It is given by ratio of photo current (I_{ph}) to incident photon power (P_0).

$$Responsivity = \frac{I_{ph}}{P_0} A/W$$

Specific detectivity (D^*): It is another critical figure of merits for photodetectors. The device with larger specific detectivity can detect weak signals, which is comparable to the device noise [20] .

$$D^* = R\sqrt{\frac{A}{2eI_d}}$$

External quantum efficiency (EQE): It is defined as the ratio of photogenerated charge carriers to the number of incident photon [21].

$$EQE = \frac{hcR}{e\lambda}$$

where h, c, R, e and λ are plank's constant, velocity of light, photoresponsivity, charge on the electron and wavelength of incident light, respectively.

Linear dynamic range (LDR): It is defined as the range of device response where the photo current is linear to incident optical power [22].

$$LDR = 20\ log\left(\frac{I^*_{ph}}{I_d}\right) dB$$

where I^*_{ph} is the photo current at an incident power of 1 mW/cm^2.

Fig. 4(a) has shown the photodetector device schematic. The active device area was 2 mm $\times$ 50 μm. As shown in fig. 4(b), the pristine ZnO based samples have shown the dark current and photo current of 0.19 μA and 3.2 μA, respectively, at an applied bias of 20 V. The photosensitivity of the pristine ZnO based device was around 15.84. Therefore, considering the lower photo response of pristine ZnO samples,

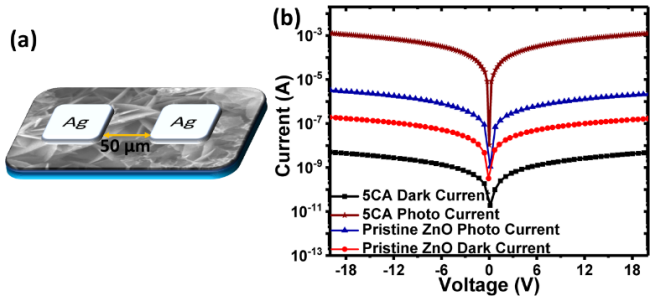


Fig. 4. (a) The device schematic and (b) photo-response of pristine ZnO nanorods and 5CA samples.

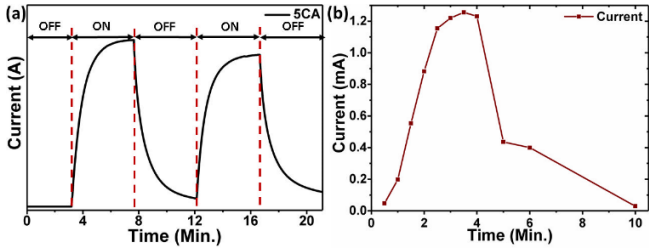


Fig. 5. (a) Switching response of the 5CA sample and (b) Change in photocurrent with respect to time.

5CA samples were considered for further I-V characterization due to their novel morphology and above-discussed properties. The dark current of the 5CA sample at 20 V applied bias was 5.17 nA. The Schottky junction at Ag/ZnO interface could be assigned to the such low dark current. Further, as the device was shined by 254 nm, 2.11 mW/cm^2 light source, the photocurrent reaches to 1.26 mA. The photo-sensitivity of the device was 2.4 $\times$ 10^5. The responsivity and specific detectivity of the device were 597 A/W and 4.64 $\times$ 10^{14} cm$\cdot$Hz$^{1/2}\cdot$W^{-1}, respectively. The EQE of the device was calculated to be 2964%. The significant improvement in the device performance can be assigned to larger surface area and high charge transport efficiency of the nanostructures [16]. Further, the linear dynamic range of the device was 65.82 dB.

Further, switching response of the 5CA sample has been recorded to investigate the response time of the device (shown in Fig. 5(a)). The rise time and fall time of the device was 107 sec. and 160 sec. respectively. Notably, a significant decrease in photocurrent after each switching cycle has been observed. Moreover, as shown in fig. 5(b), when the sample was continuously illuminated for 10 min., the photocurrent substantially increases with time till three minutes and thirty seconds, after that current suddenly start to decrease with increase in time and after 10 minutes the photocurrent value reaches to its initial value.

In order to investigate the above-observed phenomenon in more detail, PL analysis of the 5CA sample under continuous 325 nm laser illumination has been performed and data has been recorded for different duration of continuous UV illumination (Fig. 6(a)). The PL spectra taken just after UV exposure has shown a very low-intensity DLE peak. However, as exposure time increases the intensity of the DLE peak increases, which clearly suggested the formation of Oxygen vacancies related defect states over the ZnO nanostructures

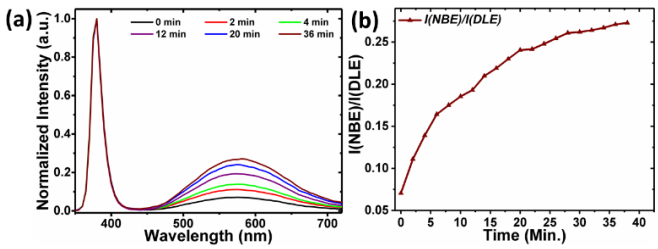


Fig. 6. (a) Change in 5CA sample PL spectra with time and (b) change in I_{NBE} to I_{DLE} ration with respect to time.

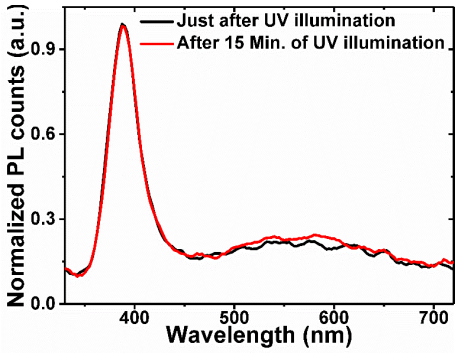


Fig. 7. PL spectra of the Pt/5CA sample just after and after 15 min. of continuous UV illumination.

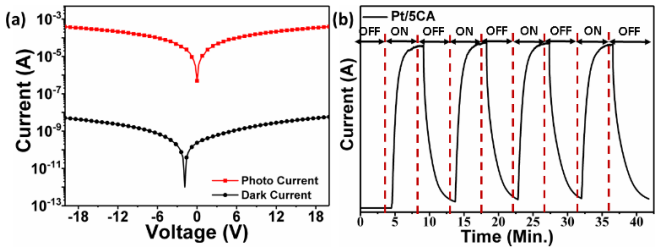


Fig. 8. (a) I-V response and (b) Switching response of the Pt/5CA device.

surface due to UV illumination [23]. Fig. 6(b) has shown the change in NBE to DLE peak intensity ratio. Initially it was 0.07 that has been increased to 0.27 after 36 min. of UV exposure. The graph shows that initially the rate of formation of Oxygen vacancies was much higher and after nearly 6 minutes, it starts to decrease. After 36 min. of continuous illumination, no significant change in NBE to DLE intensity ratio has been observed.

In order to eliminate the effect of UV exposure and to stabilize the device performance, the samples were coated with Pt nanoparticles (NPs). As shown in fig. 7, Pt coated 5CA (Pt/5CA) sample has shown no significant change in the emission spectra even after 15 min. of continuous exposure to UV radiations which shows that Pt nanoparticles have restricted the formation of Oxygen vacancies, making the device stable in UV radiations.

After confirming the stability by PL analysis in the UV radiations, I-V response of the device at 254 nm wavelength, 2.11 mW/cm^2 power light source has been taken. The device dark current and photocurrent, at 20 V applied bias was 5.3 nA and 3.13 $\times$ 10^{-4} A, respectively. The calculated device

TABLE I
COMPARISON WITH RECENTLY REPORTED UV PHOTODETECTORS

Device configuration	Sensitivity	Responsivity (AW^{-1})	Specific Detectivity ($cmHz^{0.5}W^{-1}$)	Reference
Zn_2GeO_4 nanowire	10	-	-	[14]
Au NPs coated p-Si/n-ZnO	-	3.17	-	[24]
Graphene quantum dots	~5882	2.10×10^{-3}	9.59×10^{11}	[3]
Au NPs coated ZnO	-	0.16	-	[25]
Hybrid organic photodetector	2.1×10^3	263	4.8×10^{14}	[12]
β -Ga_2O_3 nanoflakes	2.1×10^4	1403	5.99×10^{12}	[6]
$ZnGa_2O_4$ thin film	-	2782	2.29×10^{11}	[13]
MBE grown β -Ga_2O_3	1000	1.4×10^{-3}	2×10^{12}	[5]
ZnO quantum dots	1160	504	-	[1]
ZnO nanowalls	2.4×10^5	597	4.64×10^{14}	Our work

photo-sensitivity and photo-responsivity were 5.9 $\times$ 10^4 and 148 A/W, respectively. Moreover, the switching response under 254 nm light was recorded for several switching cycles. The plot has not shown any significant variation in the peak photo-current even after several switching cycles, that has also confirmed the device stability in UV radiations. Further, a slight improvement in the device response speed has been observed. Comparing to bare 5CA samples, Pt NPs coated samples have shown a decrease in rise time from 107 sec. to 90 sec., whereas the fall time remains 160 sec.

Table I shows a comparison between the recently reported UV photodetectors. the comparison was done on the basis of device figure of merits viz. photosensitivity, responsivity and specific detectivity. From the table, it was clear that, device reported in this article is better than the previously reported UV photodetectors.

IV. CONCLUSION

In summary, we have developed a high sensitivity deep UV photodetector using hydrothermally synthesized ZnO honeycomb nanostructures. The device photo-sensitivity and photo-responsivity were 2.4 $\times$ 10^5 and 597 A/W respectively. Moreover, the impact of UV radiations over device performance has been investigated and a significant decrease in device performance was observed. Therefore, to rectify this stability issue the samples were coated with Pt NPs which resulted in improved device stability and noteworthy improvement in the device response speed. The as-prepared deep UV photodetector could be a promising contender for next-generation optoelectronic devices.

ACKNOWLEDGMENT

J. Agrawal would like to thank the Sophisticated Instrument Centre, IIT Indore for providing required facilities.

REFERENCES

[1] X. Xu, C. Xu, and J. Hu, "High-performance deep ultraviolet photodetectors based on ZnO quantum dot assemblies," *J. Appl. Phys.*, vol. 116, no. 10, Sep. 2014, Art. no. 103105.
[2] A. Aldalbahi and P. Feng, "Development of 2-D boron nitride nanosheets UV photoconductive detectors," *IEEE Trans. Electron Devices*, vol. 62, no. 6, pp. 1885–1890, Jun. 2015.

[3] Q. Zhang *et al.*, "Solution-processed graphene quantum dot deep-UV photodetectors," *ACS Nano*, vol. 9, no. 2, pp. 1561–1570, Feb. 2015.

[4] J. Xu, W. Zheng, and F. Huang, "Gallium oxide solar-blind ultraviolet photodetectors: A review," *J. Mater. Chem. C*, vol. 7, no. 29, pp. 8753–8770, Jun. 2019.

[5] A. S. Pratiyush *et al.*, "Demonstration of zero bias responsivity in MBE grown β-Ga_2O_3 lateral deep-UV photodetector," *Jpn. J. Appl. Phys.*, vol. 57, no. 6, May 2018, Art. no. 060313.

[6] A. Kumar and A. Bag, "High responsivity of quasi-2D electrospun β-Ga_2O_3-based deep-UV photodetectors," *IEEE Photon. Technol. Lett.*, vol. 31, no. 8, pp. 619–622, Apr. 15, 2019.

[7] M. Sajjad, W. M. Jadwisienczak, and P. Feng, "Nanoscale structure study of boron nitride nanosheets and development of a deep-UV photodetector," *Nanoscale*, vol. 6, no. 9, pp. 4577–4582, Jan. 2014.

[8] J. Agrawal, T. Dixit, I. A. Palani, and V. Singh, "Development of Al doped ZnO nanowalls based flexible, ultralow voltage UV photodetector," *IEEE Sensors Lett.*, vol. 3, no. 9, Sep. 2019, Art. no. 3501504.

[9] B. D. Boruah, "Zinc oxide ultraviolet photodetectors: Rapid progress from conventional to self-powered photodetectors," *Nanosc. Adv.*, vol. 1, no. 6, pp. 2059–2085, Apr. 2019.

[10] L. Su, H. Chen, X. Xu, and X. Fang, "Novel BeZnO based self-powered dual-color UV photodetector realized via a one-step fabrication method," *Laser Photon. Rev.*, vol. 11, no. 6, Nov. 2017, Art. no. 1700222.

[11] C. Xie *et al.*, "Recent progress in solar-blind deep-ultraviolet photodetectors based on inorganic ultrawide bandgap semiconductors," *Adv. Funct. Mater.*, vol. 29, no. 9, Feb. 2019, Art. no. 1806006.

[12] S. Maity and T. Thomas, "Hybrid-organic-photodetector containing chemically treated ZnMgO layer with promising and reliable detectivity, responsivity and low dark current," *IEEE Trans. Device Mater. Rel.*, vol. 19, no. 1, pp. 193–200, Mar. 2019.

[13] S.-H. Tsai, S. Basu, C.-Y. Huang, L.-C. Hsu, Y.-G. Lin, and R.-H. Horng, "Deep-ultraviolet photodetectors based on epitaxial $ZnGa_2O_4$ thin films," *Sci. Rep.*, vol. 8, no. 1, Sep. 2018, Art. no. 14056.

[14] C. Yan, N. Singh, and P. S. Lee, "Wide-bandgap Zn_2GeO_4 nanowire networks as efficient ultraviolet photodetectors with fast response and recovery time," *Appl. Phys. Lett.*, vol. 96, no. 5, Feb. 2010, Art. no. 053108.

[15] R. Cuscó *et al.*, "Temperature dependence of Raman scattering in ZnO," *Phys. Rev. B, Condens. Matter*, vol. 75, no. 16, Nov. 2007, Art. no. 165202.

[16] J. Agrawal, T. Dixit, A. I. Palani, M. S. R. Rao, and V. Singh, "Zinc interstitial rich ZnO honeycomb nanostructures for deep UV photodetection," *Phys. Status Solidi Rapid Res. Lett.*, vol. 12, no. 10, Oct. 2018, Art. no. 1800241.

[17] H. Zeng, W. Cai, Y. Li, J. Hu, and P. Liu, "Composition/structural evolution and optical properties of ZnO/Zn nanoparticles by laser ablation in liquid media," *J. Phys. Chem. B*, vol. 109, no. 39, pp. 18260–18266, Sep. 2005.

[18] K. Park, J.-S. Lee, M.-Y. Sung, and S. Kim, "Structural and optical properties of ZnO nanowires synthesized from ball-milled ZnO powders," *Jpn. J. Appl. Phys.*, vol. 41, no. 12, pp. 7317–7321, Dec. 2002.

[19] J. Agrawal, T. Dixit, I. A. Palani, M. S. R. Rao, and V. Singh, "Fabrication of high responsivity deep UV photo-detector based on Na doped ZnO nanocolumns," *J. Phys. D, Appl. Phys.*, vol. 51, no. 18, Apr. 2018, Art. no. 185106.

[20] B. Hanna, K. P. Surendran, and K. N. N. Unni, "Low temperature-processed ZnO thin films for p–n junction-based visible-blind ultraviolet photodetectors," *RSC Adv.*, vol. 8, no. 65, pp. 37365–37374, Oct. 2018.

[21] J. Yao, Z. Zheng, and G. Yang, "Promoting the performance of layered-material photodetectors by alloy engineering," *ACS Appl. Mater. Interfaces*, vol. 8, no. 20, pp. 12915–12924, May 2016.

[22] L. Dou *et al.*, "Solution-processed hybrid perovskite photodetectors with high detectivity," *Nature Commun.*, vol. 5, Nov. 2014, Art. no. 5404.

[23] J. Agrawal, T. Dixit, I. A. Palani, and V. Singh, "Systematic investigations on the effect of prolong UV illumination on optoelectronic properties of ZnO honeycomb nanostructures," *Scripta Mater.*, vol. 163, pp. 1–4, Apr. 2019.

[24] J.-D. Hwang, M. J. Lai, H. Z. Chen, and M. C. Kao, "Au-mediated surface plasmon enhanced ultraviolet response of p-Si/n-ZnO nanorods photodetectors," *IEEE Photon. Technol. Lett.*, vol. 26, no. 10, pp. 1023–1026, May 15, 2014.

[25] J. D. Hwang, F. H. Wang, C. Y. Kung, and M. C. Chan, "Using the surface plasmon resonance of Au nanoparticles to enhance ultraviolet response of ZnO nanorods-based Schottky-barrier photodetectors," *IEEE Trans. Nanotechnol.*, vol. 14, no. 2, pp. 318–321, Mar. 2015.



Jitesh Agrawal was born in Nagda, India, in 1989. He received the Bachelor of Engineering degree in electronics and communication engineering from the Mahakal Institute of Technology, Ujjain, India, in 2010, and the Master of Engineering degree in embedded system and VLSI design from the Shri Vaishnav Institute of Technology, Indore, India, in 2013. He is currently pursuing the Ph.D. degree with the Indian Institute of Technology Indore, India, working on metal-oxide-based photodetector and holds Senior Research Fellowship (SRF) from University Grants Commission, New Delhi, India.



Tejendra Dixit was born in Gorakhpur, India. He received the Bachelor of Engineering degree in electronics and communication engineering from the Ajay Kumar Garg Engineering College, Ghaziabad, India, in 2012, and the Ph.D. degree from the Indian Institute of Technology Indore, India. He has worked on metal-oxide-based photodetector and also studied plasmonic behavior of noble metal nanoparticles. He is currently working in the area of 2D semiconductors as a Post-Doctoral Fellow with the Department of Physics, IIT Madras. He has published around 32 research articles in peer-reviewed journals.



I. A. Palani received the Bachelor of Engineering degree in production engineering from the Sathyabama Engineering College, Univeristy of Madras, in 2004, the Master of Engineering degree in manufacturing engineering from the A.C. College of Engineering, Anna University, in 2006, and the Ph.D. degree from the Precision Engineering Laboratory, Department of Mechanical Engineering, IIT Madras, in 2010. He was a Post-Doctoral Research Scientist with Kyushu University, Fukuoka, Japan, from September 2010 to December 2011. He has joined IIT Indore as an Assistant Professor, in January 2012. He is currently an Associate Professor with the Discipline of Mechanical Engineering, IIT Indore, India. His area of research interests are smart materials and structures, shape memory alloy, laser assisted surface processing, laser-based micromanufacturing, mechatronics system design, and optical Instrumentation.



Vipul Singh was born in Siwan, Bihar, India, in 1979. He received the bachelor's degree in electronics and communication engineering from the Delhi College of Engineering (under University of Delhi) in 2003, the M.S. degree from the Kyushu Institute of Technology, Japan, and the Ph.D. degree in March 2009. He further continued his doctoral research work under Prof. K. Kaneto of the Department of Biological Functions and Engineering, Kyushu Institute of Technology, Japan. Most of his doctoral work was based on study of interface related effects in conducting polymer-based organic optoelectronic devices. From April 2009 to November 2010, he was a Scientific Researcher with the Nanosystems Integration Laboratory, Research Institute of Electronics, Shizuoka University, Japan, where he studied MOSFET-based Single electron Transistors/ devices. Since then, he has been interested in Bulk/SOI technologies and has also studied and reduced low frequency noise occurring in these devices due to interface traps both at room and cryogenic temperatures. Since November 2010, he has also been an Assistant Professor with the Department of Electrical Engineering, Indian Institute of Technology (IIT) Indore. His interest in the area of organic electronics ranges from device applications viz. organic field effect transistors (OFETs), organic solar cells (OSCs), and organic light emitting diodes (OLEDs) to the study of charge carrier transport or photo carrier generation in various organic materials.

Pollution, 6(1): 135-147, Winter 2020
DOI: 10.22059/poll.2019.284612.642
Print ISSN: 2383-451X Online ISSN: 2383-4501
Web Page: https://jpoll.ut.ac.ir, Email: jpoll@ut.ac.ir

Development of the Ethyl Ester from Jatropa Oil through Response Surface Methodology Approach

Gautam, R.[1], Ansari, N.[1], Sharma, A.[2] and Singh, Y.[3*]

1. Department of Mechanical Engineering, Delhi Technological University, Delhi, India
2. Department of Mechanical Engineering, G L Bajaj Institute of Technology and Management, Greater Noida, UP, India
3. Department of Mechanical Engineering, Sir Padampat Singhania University, Udaipur, Rajasthan, India

Received: 02.07.2019 Accepted: 10.11.2019

ABSTRACT: With an increase in the global pollution, there is requirement for an alternative to the fossil fuels. Non-edible vegetable oils are highly promising for producing liquid fuels like diesel. Jatropha is a potential feedstock for biodiesel, currently utilized in India and many parts of the world. The optimization of reaction conditions such as temperature, time, catalyst and molar ratio for biodiesel production is important in reactor design. However, oils have characteristics reaction properties for optimum yield. Therefore, there is the need to identify such parameters in Jatropha oil ethyl esters production. Preparation of biodiesel from Jatropha oil ethyl esters using conventional homogeneous process. Optimization of Jatropha ethyl esters using Response surface methodology is done and data so obtained are fed to the design experiment software for analysis. The Jatropha ethyl esters yield was 92.62%. Maximum production of Jatropha oil ethyl ester was achieved with the process parameters viz molar ratio 8.5, reaction time 89.67min, reaction temperature 70.1°C and catalyst.0.62wt%.

Keywords: Jatropha oil; biodiesel production; reaction conditions; biodiesel oil.

INTRODUCTION

Energy has a vital part in the progress of a nation and is measured in terms of Economic and social development of that country (Chauhan et al., 2010; Chauhan et al., 2012; Chauhan et al., 2016). Measure of prosperity of a country depends on per capita energy consumption besides per capita income and GDP. Fossil fuels (petroleum oil, coal, and natural gas), renewable energy (hydro, wind, solar, geothermal, marine energy, and combustible wastes), and nuclear energy are the major source of world energy consumption (Ansari et al., 2018; Sharma et al., 2019; Singh et al., 2018). Petroleum is the largest pool of energy consumption by the world. The ever-rising demand of transportation fuel has been fulfilled by Petro based fuels (approx. ~ 95%) (Cho & Chauhan, 2009). Processed vegetable oils (biodiesel) both edible & nonedible are being considered as a promising alternative to diesel for its use in a diesel engine (Chauhan et al., 2010, Chauhan et al., 2016, Cho & Chauhan, 2009; Chauhan et al. 2009; Rai et al., 2013; Singh et al., 2017). The impact of

* Corresponding Author, Email: yashvirkankara21@gmail.com

large-scale production of biodiesel and their usage are perceptible and a paradigm shift in consumption pattern has been visualized, and this is clearly encouraging the demand for biofuels (Singh et al., 2017; Singh et al., 2019).

The very fact that fossil petroleum diesel is limited and its long-term availability is required and the concern about the environmental norms and increasing pollution level forced to search the renewable alternatives in order to rectify current problems (Acharya et al., 2017; Agarwal et al., 2008; Agarwal et al., 2006; Algayyim et al., 2018). In recent decades, a balanced approach has been made by the scientists and researchers to increase the use of non- edible vegetable oil and biodiesel to draw a common policy nation wise to implement as genuine substitute to fossil diesel in engines without much modification the existing hardware (Ganesan et al., 2018; Sakthivel et al., 2013). Fuels from bio origins are also the solution of energy security, employment, restricting the country wealth used in importing fossil fuels, increasing agriculture economy, using waste land to grow oil rich crops and importantly environment pollution and degradation (Gautam & Kumar 2015; Gautam & Kumar 2018). A 10% biodiesel blend shows the best engine performance in terms of engine torque, engine power, fuel consumption and brake thermal efficiency among the all blending ratios for the three biodiesel blends (Ong et al., 2014). Blending J50C50 biodiesel with diesel reduces the carbon dioxide emissions and smoke opacity, but increases the nitrogen oxide and carbon monoxide emissions (Dharma et al., 2017). Literature review also confirms that methyl ester of Jatropha oil (JOME) has been extensively studied as a fuel with various approaches but ethyl ester of Jatropha oil (JOEE) as a fuel need to study because both the ingredients are renewable and can be easily derived from biomass (Singh et al., 2012).

Optimization of the yield of Papaya oil methyl ester was investigated using response surface methodology. Within the range of the selected operating conditions, the optimized values of temperature, catalyst amount, time, and methanol to oil molar ratio were found to be 62.33□°C, 0.95□wt %, 3.30□min, and 9.50:1 respectively. Based on the optimum condition, the predicted biodiesel yield was 99.9% and the actual experimental value was 99.3%. Papaya oil methyl ester (POME) exhibits property close to ASTM standards (Nayak et al., 2019). The optimum combinations for the reaction were CALB:RML ratio (2.5:1), t-butanol to oil (39.9□wt%), temperature (35.6□°C), methanol:oil ratio (5.9), reaction time 33.5□h. FAME yield of 78.3.5%, which was very close to the predicted value of 75.2%, was obtained (Shahedi et al., 2019). In the present study the different mechanism of biodiesel production is carried out with two stage integrated acid-catalyzed and pre-esterification of FFA and base-catalysed transeterification process to find the best techniques for production of fine biodiesel.

MATERIAL AND METHODS

Through response surface methodology (RSM) mathematical and statistical techniques are taken for analyzing problems by the influence of several independent variables and dependent variable or response, and the goal is to optimize this response (Tacias-Pascacio et al., 2019).

For response surface methodology, the dependent variable was viewed for the surface to which a mathematical representation was fitted. The regression equations were developed, related to various quality characteristics of biodiesel production, the second order response surface was assumed, as given in eq. (1):

$$Y = b_0 + \sum_{i=1}^{k} b_i x_i + \sum_{i=1}^{k} b_{ii} x_i^2 + \sum_{i<j=2}^{2} b_{ij} x_i x_j \pm e_r \quad (1)$$

The surface Y contains linear, squared and cross product terms of variables x_i‟ s. It has been proposed that the scheme based on central composite rotatable design fits the second-order response surfaces quite accurately (Chauhan et al., 2016, Ansari et al., 2018, Sharma et al. 2018).

In this design, the standard error remains the same at all the points which are middle from the center of the region. This criterion of rotatability could be explained as follows: Let the point (0, 0, ---, 0) symbolize the center of the region in which the relation between Y and X is under investigation. From the results of any experiment, the standard error, e_r of Y can be computed at any point on the fitted surface. This standard error acts as a meaning of the co-ordinates x_i‟ s of the selected point (Atapour et al. 2014; Awad et al., 2017; Hirkude & Padalkar, 2014). Thus, because of rotatability form, the standard error remains same at all equidistant points with the distance ρ from center of region i.e. for all points, this is given by the equation (2).

$$X_1^2 + X_2^2 + \ldots\ldots\ldots\ldots + X_k^2 = \rho^2 = \text{constant} \quad (2)$$

Central composite rotatable design is divided in three parts as-

1. Points related to 2^k design, where 2 is the number of levels where the parameters are kept during testing and k is the number of parameters
2. Extra points (star points), positioned on the co-ordinate's axes for a central composite design with star arm of size α
3. Additional points added at the center for equal precision to response Y of circle with radius one

Factor α gives the radius of circle or sphere on which the star points lie. For k ≥5, experimental size is reduced for half replication of 2^k factorial design. Thus, α become $2^{(k-1)/4}$ for half replication. Also, no duplication is needed to find error mean square by replicating the center points (Bharawaz et al., 2016; Daud et al., 2018; Gopal et al 2018). In Table 1, components of second order middle composite rotatable design for a different number of variables are given.

Table 1. Components of central composite second order rotatable design

Variables (k)	Factorial Point (2^k)	Star Point (2k)	Center Points (n)	Total (N)	Value of α
3	8	6	6	20	1.682
4	8	8	5	21	1.672
5	16	10	6	32	2.000
6	32	12	9	53	2.378

The regression equation demonstrating second order response surface given by Eq. 3:

$$Y = b_0 + \sum_{i=1}^{k} b_i x_i + \sum_{i=1}^{k} b_{ii} x_i^2 + \sum_{i<j=2}^{2} b_{ij} x_i x_j \pm e_r \quad (3)$$

where, Y = estimated response, b‟ s are the coefficients and x_i‟ s are the independent variables.

As intended, both stages of transesterification i.e. acid catalyzed and base catalyzed would be optimized. The independent variables were chosen from the preliminary studies conducted earlier which identified the most important factors affecting the esterification/ transesterification reactions of the Jatropha vegetable oils. For the esterification stage of methyl and ethyl both, concentration of catalyst (Para-toluene sulfonic acid-PTSA in this case), molar ratio, catalytic concentration, reaction. Time and reaction. Temperature were. Considered as the critical process. parameters, whereas for the alkaline stage, concentration of potassium hydroxide (KOH), reaction time and reaction. Temperature were considered.

The design variables and levels are presented in table 2 for the esterification stage of Jatropha ethyl esters.

Table 2. Process Parameters and Their Levels For Esterification Jatropha Ethyl Esters

Coded Factors	Real factors	Parameters	1		Levels		
			(-1.682)	(-1)	(0)	(+1)	(+1.682)
X_1	A	Molar Ratio	0.6	3.5	7.75	12	14.89
X_2	B	Catalyst Concentration (.% wt.)	.0.16	.0.5	.1	.1.5	.1.84
X_3	C	Reaction temperature (°C)	28.06	40	57.5	75	86.93
X_4	D	Reaction time (minute)	9.54	30	60	90	110.45

RESULTS AND DISCUSSIONS

The data so obtained were fed to the design experiment software for further analysis. The analysis of variance (ANOVA) and the statistical values were obtained from the software as shown in table 3. and table 4. The optimal FFA was .4.74% from the experimental design study. However, studies show that it should be less than.2% for higher yield. Therefore, esterification was done using the optimal parameters to obtain.4.74% FFA in.50 minutes. After that extra .20 minutes agitation given to get the requisite amount of FFA (<.2%). Moreover, the oil was transesterified as per the optimal process parameters of .0.88% of KOH by original oil mass and approx. 60°C temperature for.70 minutes. The sample from the reactor was gravity separated for.12 hours to remove glycerol. For purification, the sample was washed many times with lukewarm distilled water to remove traces of methanol/ethanol or catalysts that remained. After, the oil heated at.100°C for.30 minutes to evaporate the moisture contents. Lastly, Jatropha ethyl esters obtained as a light yellow and see through fluid. The. reaction. time, reaction. Temperature, molar. ratio and concentration of catalyst poly toluene sulphonic acid (.PTSA) are considered as the factors and the. % FFA as the response in the esterification stage.

Same procedure was followed as like esterification of methyl ester whereas the representation's F-value of.7.57 implied the representation is important. Probability of occurrence of such a F-value because of noise is very low about 0.14%. The terms in the representation would be important if "Prob.> F" is less than.0.0500. In this case, B, .C, .AC, .AD, .BD, $.C^2$ are important representation's terms. For values more than 1000, the representation terms are not important. "Lack of Fit F-value" of .2.73 implied that the Lack of F is of no significance, relative to pure error. There is 17.47% probability that this huge "Lack of Fit F-value" could be present due to noise. It is known that a non-important lack of fit is desirable, so it is concluded that the representation fitted the design. ."Pred R-Squared" of.0.4153 is not very close to "Adj R-Squared" of.0.7472 as expected; i.e. the difference was greater than.0.2. This indicates a large-block effect or probably a problem with your representation and/or data. Things like a representation's reduction, response transformation, outliners are to be taken into account. "Adequate Precision" measured the signal to noise ratio. The required ratio was a value more than 4. So, a ratio of 10.475 indicated an adequate signal. These representations can help to steer through the design space.

Equation of coded factors make predictions about response for given levels of each factor with +1 for high levels of the factors and –1 for lower levels. Relative impact of factors can be calculated by comparing factor coefficients with the help of coded equation.

The prediction of response for the levels of each factor can be done by applying the equation using actual factors. Here, levels have been specified in the original units for.each factor. This equation could not.be used to determine relative impact of each factor, because the coefficients are

multiplied in order to keep the units of each factor in check and the intercept is obtained away from the center of the design space. As, all representation's statistics and diagnostic plotted are important, the Representation's Graphs icons are finished up. Statistical considerations are satisfied for-

1) Normal probability plot to find the normality of standardized residuals.

2) Standardized residuals versus estimated values: Check for constant error.

3) To monitor influential values for externally standardized residuals

4) For power transformations using box-cox plot.

Table 3. Analysis. of Variance for. Esterification

ANOVA. for Response Surface Reduced Quadratic representation's						
Analysis. of variance [Partial sum of squares]						
Source	**Sum of Squares**	**df**	**Mean Square**	**F-Value**	**p-value Prob> F**	**Remarks**
Representation's	52.60	9	5.84	7.57	0.0014	Significant
A-Molar ratio	2.00	1	2.00	2.59	0.1359	
B-Catalyst concentration	11.52	1	11.52	14.92	0.0026	
C-Reaction temperature	4.56	1	4.56	5.91	0.0333	
D-Reaction time	1.41	1	1.41	1.83	0.2035	
AC	7.03	1	7.03	9.10	0.0117	
AD	8.64	1	8.64	11.19	0.0065	
BD	9.10	1	9.10	11.78	0.0056	
C^2	16.72	1	16.72	21.66	0.0007	
D^2	1.680E-003	1	1.680E-003	2.176E-003	0.9636	
Residual	8.50	11	0.77			
Lack of Fit	7.02	7	1.00	2.73	0.1747	Non-Significant
Pure Error	1.47	4	0.37			
Total	61.10	20				

Table 4. Statistical values of the Ethyl Esterification of Jatropha oil

Std. Dev.	0.88	R-Squared	0.8610
Mean.	4.58	Adj R-Squared	0.7472
C.V. %	19.20	Pred R-Squared	0.4153
PRESS.	35.73	Adeq Precision	10.475
-2 Log Likelihood.	40.59	BIC	71.03
		AICc	82.59

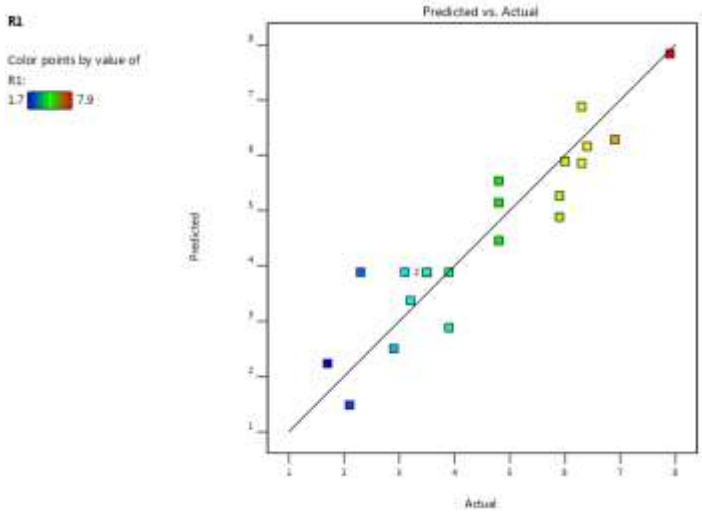


Fig. 1. Estimated and Actual Values for Esterification Jatropha Ethyl Ester

Fig.1 shows the estimated and the actual values of the design. It may be clearly seen that the estimated and actual values grouped are quite similar and concentrating on a medium range. In the light of the above discussions, it may be stated that the representations had effectively estimated the %FFA for large sized esterification with minimal error. It was actual that the estimated and the actual %FFA are very close in all the cases validating the previous assumption that the representation was statistically important. Now the three-dimensional surfaces were plotted using the representation in design experiment software. Response surface graphs represent the main and.interactive.effects of the factors on the response (% FFA). Results so obtained are shown in Fig. 2, Fig. 3 and Fig. 4, respectively.

Relationships between the parameters are indicated by the two independent variables that are plotted and converted on three-dimensional contour lines. In Figure 2, the change in the conversion of FFA having a molar ratio of ethanol.to acidified oil with reaction temperature are shown. It is seen that at high reaction temperature, increase in molar ratio leads to increase in conversion of FFA and at high molar ratio decrease in reaction temperature led to decrease in FFA conversion. If biodiesel is produced using a solid catalyst, reaction kinetics show that the reaction depends on the temperature at one stage and the results are in line with similar type of work (Anwar et al., 2018; Elango et al., 2019).

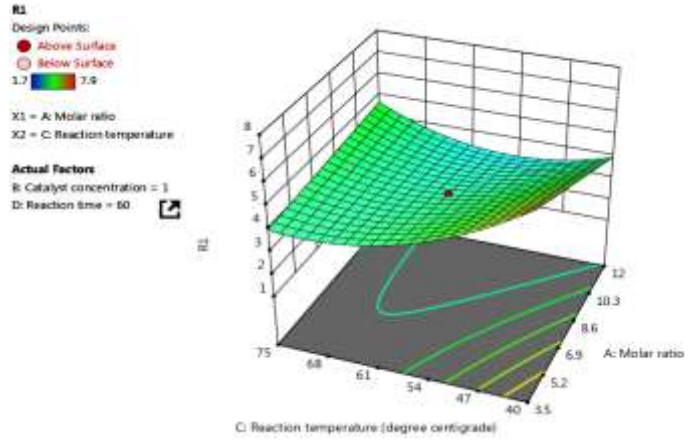


Fig. 2. Response Plot for Effect of Reaction Temperature and Molar Ratio on Esterification of Ethyl Ester

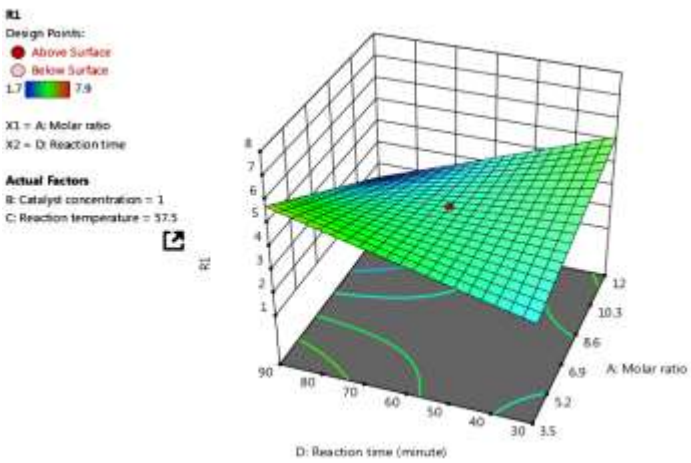


Fig. 3. Response Plot for Effect of Reaction Time and Molar Ratio on Esterification of Ethyl Ester

Figure.3 shows that the change of FFA attains a peak value.at low molar ratios of ethanol to acidified oil and longer reaction time. Increasing the molar ratio of ethanol to acidified oil does not improve the conversion of FFA due to decreased concentration of active catalyst sites in the reaction (Dharma et al., 2016, Atapour et al., 2014). Also, hydrogen bonds between ethanol and sulfonic group ($-SO_3H$) of the catalyst form easily due to the polarity of ethanol and ethanol molecules adsorbed on active sites of the catalyst. Concentration of ethanol reached a certain level, then due to the poor accessibility of catalyst surface, esterification was stopped. Therefore, high conversions were obtained at long reaction time and low molar ratios of ethanol to acidified oil.

Fig. 4 presents the response plot for reaction time and catalyst loading with respect to the conversion of FFA. Thus, long reaction time and high amount of catalyst aided high conversion of FFA. Also, with reduced catalyst action and reaction times, conversion of FFA decreased. This occurs as the percentage of catalyst loading reduces which decreases the number of reactants adsorbed on the active sites of the catalyst. And the conversion of FFA increases with increase in the reaction time (Esonye et al., 2019; Karmakar et al., 2018). Optimum FFA of 1.9 % was given by the representation and catalysts concentration of 1.0%, with reaction temperature equal to 59°C and reaction time of 72 minutes. Thus, the catalyst concentration reduced the FFA value below 2%, which is recommended for transesterification. However, looking at the initial FFA of 13.7 % the reduction was found to be encouraging.

The table 5 shows the analysis of variance for the transesterification of ethyl ester. The Representation's F-value of 592.87 implied the representation held importance. An F-value being large due to noise was unlikely, with just 0.01% chance. Values of Prob> F, less than 0.0500 shows that representation's terms are important. For such case, A^2, B^2, C^2, D^2, B, C, D, BC, CD, AD shows important representation's terms. For values more than 0.1000 shows that representation's terms are not important. Lack of Fit F-value as 1.18 prove that Lack of Fit was not important relative to pure error. 44.89% of Lack of Fit F-value is due to noise. Lack of fit as non-important is desired (Elango et al., 2019; Dharma et al., 2016; Esonye et al., 2019). The statistical values for Transesterification are as shown by Table 6.

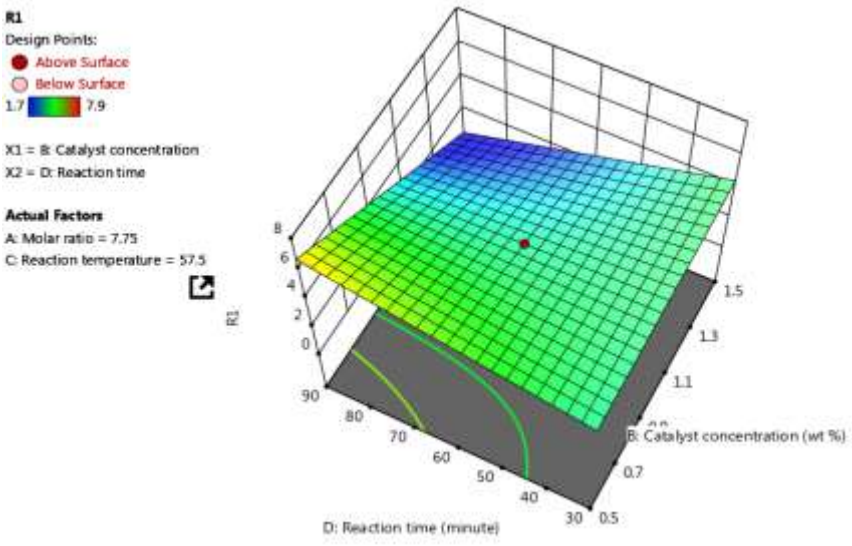


Fig. 4. Response Plot for Variation of Esterification of Ethyl Ester with Reaction Time and Catalyst Concentration

Table 5. Analysis of Variance for Transesterification of Ethyl Ester

ANOVA for Response Surface Reduced Quadratic representation						
Analysis of Variance [Partial sum of squares]						
Source	**Sum of Squares**	**df**	**Mean Square**	**F Value**	**p-value Prob> F**	**Result**
Representation	2295.53	11	208.68	592.87	< 0.0001	Significant
A-Molar ratio	0.067	1	0.067	0.19	0.6727	
B-Catalyst Concentration	25.21	1	25.21	71.61	< 0.0001	
C-Temperature	1539.15	1	1539.15	4372.70	< 0.0001	
D-Time	15.38	1	15.38	43.69	< 0.0001	
AD	36.00	1	36.00	102.26	< 0.0001	
BC	19.59	1	19.59	55.67	< 0.0001	
CD	17.64	1	17.64	50.12	< 0.0001	
A^2	11.82	1	11.82	33.59	0.0003	
B^2	15.02	1	15.02	42.68	0.0001	
C^2	659.24	1	659.24	1872.87	< 0.0001	
D^2	7.37	1	7.37	20.93	0.0013	
Residual	3.17	9	0.35			
Absence of Fit	1.89	5	0.38	1.18	0.4489	Non-Significant
Absolute Error	1.28	4	0.32			
Corr. Total	2298.70	20				

Table 6. Statistical Values for Transesterification of Ethyl Ester

Std. Dev.	**0.59**	**R-Squared**	**0.9986**
Mean	85.49	Adj R-Squared	0.9969
C.V. %	0.69	Pred R-Squared	0.9859
PRESS	32.40	Adeq Precision	86.411
-2 Log Likelihood	19.88	BIC	56.41
		AICc	82.88

Pred R-Squared of 0.9859 results in Adj R-Squared of 0.9969; i.e. the difference was less than 0.2. Signal to noise ratio was measured by Adeq Precision. Ratio more than 4 was required. Ratio of 86.411 gives desirable signal. This representation can also be used to steer through the design space.

$$\text{Ethyl Ester} = +91.50-0.070* A+2.11* B+10.62* C+1.06* D+3.30* AD+1.56* BC-1.48* CD-0.89* A^2-1.00* B^2-6.64* C^2-0.70* D^2 \quad (4)$$

Equation 4 of coded factors make predictions about response for given levels of each factor with +1 for high levels of the factors and –1 for lower levels. Relative impact of factors can be calculated by comparing factor coefficients with the help of coded equation.

$$\text{Ethyl Ester} = - 27.94301 -1.57967 * A + 24.94843 * B + 2.99756 * C + 0.16572 * D + 0.031020 * A * D +0.44714* B * C - 3.39429E -003 * C * D-0.049245* A^2-25.06558* B^2-0.021688* C^2-1.12336E-003* D^2 \quad (5)$$

Equation 5 for actual factors is for making assumptions about the response for different levels of each factor.

When statistical conditions are fulfilled, the results are shown in figure 5.

Here, levels have been specified in the original units for each factor. This equation could not.be used to determine relative impact of each factor, because the coefficients are multiplied in order to keep

the units of each factor in check and the intercept is obtained away from the center of the design space. When statistical conditions are fulfilled, the results are shown in Figure 5.

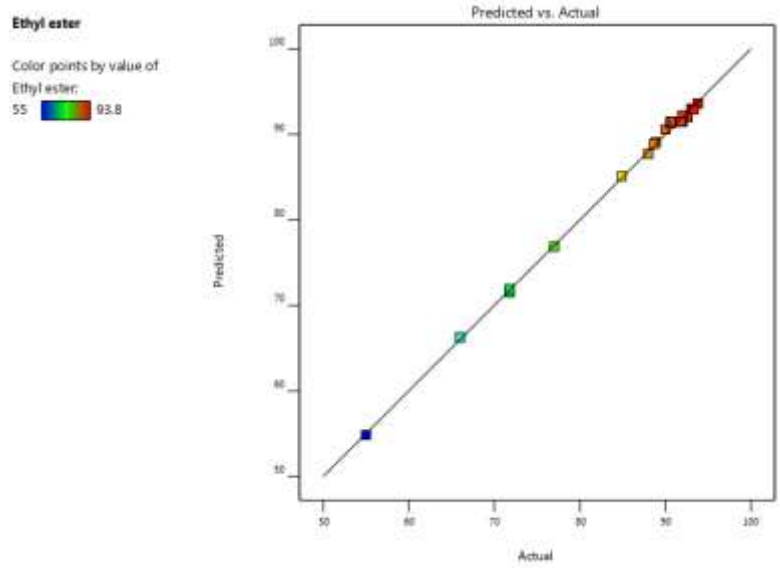


Fig. 5. Estimated and actual values Transesterification Jatropha Ethyl Ester

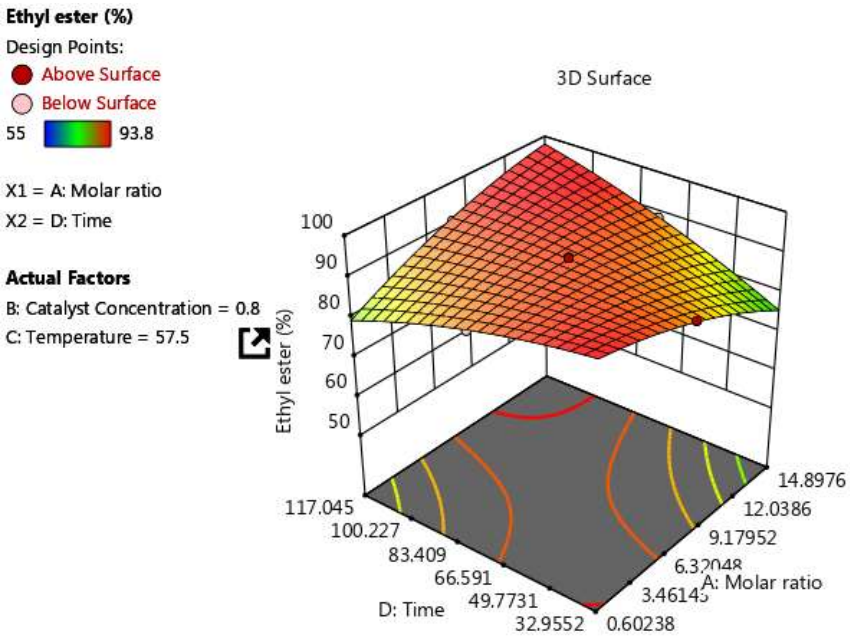


Fig. 6. Response Plot for Variation of Transesterification of Ethyl Ester with Molar Ratio and Time

Figure 6 is response surface plot of the effect of molar ratio and time with constant catalyst concentration and temperature and the influence of reaction time and quantity of ethanol on the JOEE yield. Molar ratio (ethanol/oil) is most important for the transesterification for producing biodiesel and the separation, recovery of glycerol from oil depends on it. 6:1 is considered the best molar ratio for methanol, in the preset study molar ratio of 9:1 is best for ethanol. The biodiesel yield increased at

higher time and high ethanol/oil ratio because the conversion of triglycerides increases with time and excess ethanol favored the complete conversion to biodiesel (Sharma et al., 2019, Ansari et al., 2018, Singh Y., 2015).

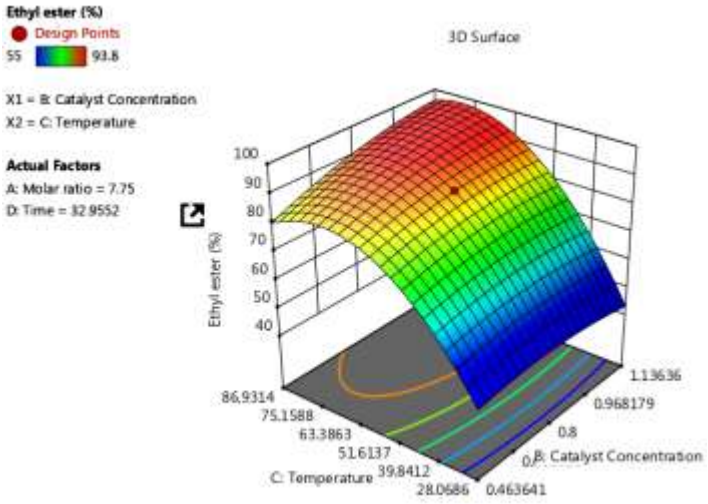


Fig. 7. Response Plot for Variation of Transesterification of Ethyl Ester with Catalyst Concentration and Temperature

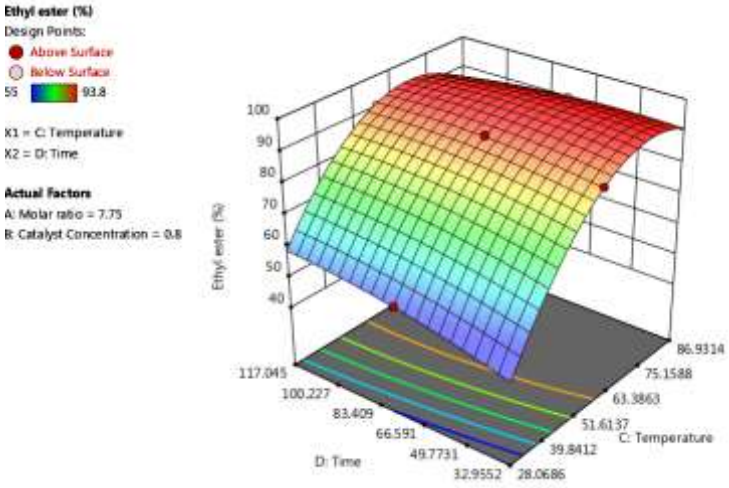


Fig. 8. Response Plot for Effect of Time and Temperature on Transesterification of Ethyl Ester

Figure 7 presents the response surface plot of the variation in biodiesel yield with temperature and catalyst concentration when the ethanol-oil ratio is 7.75 during 75 minutes. NaOH acts as catalyst at elevated temperatures as transesterification reaction is faster than saponification reaction. Thus, temperature is main factor to increase the rate of reaction for producing biodiesel. At lower molar ratio (3:1) yields increases with increase in catalyst concentration. At higher molar ratio (7.5:1), ethyl ester produced, increased. Hence, in improving the ethyl ester yields, catalyst concentration is the main variable. Biodiesel produced has problems in separation if too much catalyst is used, due to emulsion formation. Too much catalyst, over 1.0wt % decreased biodiesel production due to soap formation that inhibits ester layer formations (Sharma et al., 2019, Jamshaid et al., 2019, Yatish et al., 2018).

Figure 8 presents the response surface plot of the variation in biodiesel yield with

time and catalyst concentration when the ethanol-oil ratio was 7.75:1 at 0.8% wt. as catalyst concentration obtains the yield 93.5%. Biodiesel yield increased with higher catalyst concentration. The usage of too much catalyst led to saponification reaction and reduced biodiesel yield (Sharma et al., 2019, Atapour et al., 2014, Sathish Kumar et al., 2015). Optimum of 92.62 % was estimated by the representation's and catalysts concentration of 0.62%, molar ratio 8.5: 1, reaction temperature of 70.12°C and reaction time of 89.67 minutes for minimize the glycerol and maximize the production of ethyl ester. Also, ethanol can be used as a substitute for the methanol in optimized conditions. Thus, biodiesel can be made from non-conventional sources (like ethanol obtained from sugar cane, soybean oil). This study shows the usage of ethanol or methanol along with Jatropha oil to produce fuel.

CONCLUSIONS

The present study was done on Jatropha oil which was converted into Jatropha ethyl ester and optimization for production of Jatropha oil ethyl esters using response surface method. RSM proved to be a powerful tool for the optimization of biodiesel production. It has been reported that the yield of biodiesel was high enough to produce biodiesel in a practical way.

- Maximum production of Jatropha oil ethyl ester was achieved with the process parameters viz molar ratio 8.5, reaction time 89.67min, reaction temperature 70.1°C and catalyst.0.62wt%. An empirical relationship has been developed to predict the product yield incorporating transesterification process parameters at 92.62 %. ± 0.3 %.
- Confidence level for the studies on different mechanism of biodiesel production revealed that two stage integrated acid- catalyzed pre-esterification of FFA and base-catalyzed transeterification process may be considered the best for production of biodiesel.
- RSM proved to be a powerful tool for the optimization of biodiesel production. It has been reported that the yield of biodiesel was high enough to produce biodiesel in a practical way. The optimal conditions for the maximum methyl ester yield were found to be at catalyst concentration of 0.57% w/w, methanol/oil molar ratio of 7.16, and a reaction temperature of 59.43°C.
- An empirical relationship has been developed to predict the product yield incorporating transesterification process parameters at 95% confidence level. The predicted value for yield is found 94.48%. The error in the experimental and the statistical model results was only 2.08%

GRANT SUPPORT DETAILS

The present research did not receive any financial support.

CONFLICT OF INTEREST

The authors declare that there is not any conflict of interests regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/ or falsification, double publication and/or submission, and redundancy has been completely observed by the authors.

LIFE SCIENCE REPORTING

No life science threat was practiced in this research.

REFERENCES

Acharya, N., Nanda, P., Panda, S. and Acharya, P. (2017). A comparative study of stability characteristics of mahua and jatropha biodiesel and their blends. *Journal of King Saud University - Engineering Sciences*.

Agarwal, D., Kumar, L. and Agarwal, A. K. (2008). Performance evaluation of a vegetable oil fuelled

compression ignition engine. *Renewable Energy*, 33, 1147-1156.

Agarwal, D., Sinha, S. and Agarwal, A. K. (2006). Experimental investigation of control of NOx emissions in biodiesel-fueled compression ignition engine. *Renewable Energy*, 31, 2356-2369, 20.

Algayyim, S. J. M., Wandel, A. P. and Yusaf, T. (2018). Butanol–acetone mixture blended with cottonseed biodiesel: Spray characteristics evolution, combustion characteristics, engine performance and emission. *Proceedings of the Combustion Institute*, 37.

Ansari, N. A., Sharma, A. and Singh, Y. (2018). Performance and emission analysis of a diesel engine implementing polanga biodiesel and optimization using Taguchi method. *Process Safety and Environmental Protection*, 120, 146-154.

Anwar, M., Rasul, M. G. and Ashwath, N. (2018). Production optimization and quality assessment of papaya (Carica papaya) biodiesel with response surface methodology. *Energy Conversion and Management*, 156, 103-112.

Atapour, M., Kariminia, H.-R. and Moslehabadi, P. M. (2014). Optimization of biodiesel production by alkali-catalyzed transesterification of used frying oil. *Process Safety and Environmental Protection*, 92, 179-185.

Awad, O. I., Mamat, R., Ali, O. M. and Azmi, W. H. (2017). Response surface methodology (RSM) based multi-objective optimization of fusel oil -gasoline blends at different water content in SI engine. *Energy Conversion and Management*, 150, 222-241.

Chauhan, B.S., Kumar, N., Pal, S. S.and Du Jun, Y.(2011). Experimental studies on fumigation of ethanol in a small capacity diesel engine. *Energy*, 36, 1030-1038.

Chauhan, B.S., Kumar, N. and Cho, H. M. (2010). Performance and emission studies on an agriculture engine on neat Jatropha oil. *Journal of Mechanical Science and Technology*, 24, 529-535.

Chauhan, B.S., Kumar, N.and Cho, H. M. (2012). A study on the performance and emission of a diesel engine fueled with Jatropha biodiesel oil and its blends. *Energy*, 37, 616-622.

Chauhan, B.S., Kumar, N., Du Jun, Y. and Lee, K. B. (2010). Performance and emission study of preheated Jatropha oil on medium capacity diesel engine. *Energy*, 35, 2484-2492.

Chauhan, B.S. Singh, R. K. Cho, H. and Lim, H. (2016). Practice of diesel fuel blends using alternative fuels: a review. *Renewable and Sustainable Energy Reviews*, 59, 1358-1368.

Cho, H.M. and Chauhan, B. S. (2009). The Study on Performance and Emission of CNG as a Potential Fuel in Kore. *Journal of the Korean Institute of Gas*, 13, 39-43.

Elango, R. K., Sathiasivan, K., Muthukumaran, C., Thangavelu, V., Rajesh, M. and Tamilarasan, K. (2019). Transesterification of castor oil for biodiesel production: Process optimization and characterization. *Microchemical Journal*, 145, 1162-1168.

Datta Bharadwaz, Y., Govinda Rao, B., Dharma Rao, V., and Anusha, C. (2016). Improvement of biodiesel methanol blends performance in a variable compression ratio engine using response surface methodology. *Alexandria Engineering Journal*, 55, 1201-1209.

Daud, N. M., Sheikh Abdullah, S. R. and Abu Hasan, H. (2018). Response surface methodological analysis for the optimization of acid-catalyzed transesterification biodiesel wastewater pre-treatment using coagulation–flocculation process. *Process Safety and Environmental Protection*, 113, 184-192.

Dharma, S., Hassan, M. H., Ong, H. C., Sebayang, A.H., Silitonga, A.S., Kusumo, F. and Milano, J. (2017). Experimental study and prediction of the performance and exhaust emissions of mixed Jatropha curcas-Ceiba pentandra biodiesel blends in diesel engine using artificial neural networks. *Journal of Cleaner Production*, 164, 618-633.

Dharma, S., Masjuki, H., Ong, H. C., Sebayang, A., Silitonga, A. and Kusumo, F. (2016). Optimization of biodiesel production process for mixed Jatropha curcas–Ceiba pentandra biodiesel using response surface methodology. *Energy Conversion and Management*, 115, 178-190.

Esonye, C., Onukwuli, O. D. and Ofoefule, A. U. (2019). Optimization of methyl ester production from Prunus Amygdalus seed oil using response surface methodology and Artificial Neural Networks. *Renewable Energy*, 130, 61-72.

Ganesan, S., Senthil Kumar, J. and Hemanandh, J. (2018). Optimisation of CI engine parameter using blends of biodiesel by the Taguchi method. *International Journal of Ambient Energy*, 1-4.

Gautam, R. and Kumar, N. (2015). Comparative study of performance and emission characteristics of Jatropha alkyl ester/butanol/diesel blends in a small capacity CI engine. *Biofuels*, 6, 179-190.

Gautam, R. and Kumar, N. (2018). Performance emission and combustion studies of diesel engine on Jatropha ethyl ester and its higher alcohol blends. *International Journal of Global Warming*, 14, 159-169.

Gopal, K., Sathiyagnanam, A. P., Rajesh Kumar, B., Saravanan, S., Rana, D. and Sethuramasamyraja, B. (2018). Prediction of emissions and performance of a diesel engine fueled with n-octanol/diesel blends using response surface methodology. *Journal of Cleaner Production*, 184, 423-439.

Hirkude, J. B. and Padalkar, A. S. (2014). Performance optimization of CI engine fuelled with waste fried oil methyl ester-diesel blend using response surface methodology. *Fuel*, 119, 266-273.

Jamshaid, M., Masjuki, H. H., Kalam, M. A., Zulkifli, N. W. M., Arslan, A. and Alwi, A. (2019). Production optimization and tribological characteristics of cottonseed oil methyl ester. *Journal of Cleaner Production*, 209, 62-73.

Karmakar, B., Dhawane, S. H. and Halder, G. (2018). Optimization of biodiesel production from castor oil by Taguchi design. *Journal of Environmental Chemical Engineering*, 6, 2684-2695.

Nayak, M.G. and Vyas, A.P. (2019). Optimization of microwave-assisted biodiesel production from Papaya oil using response surface methodology. *Renewable Energy*, 138, 18-28.

Ong, H.C., Masjuki, H.H., Mahlia, T.M.I., Silitonga, A.S., Chong, W.T. and Yusaf, T. (2014). Engine perfromance and emissions using Jatropha curcas, Ceiba pentandra, and Calophyllum inophyllum biodiesel in a CI diesel engine. *Energy*, 69, 427-445.

Rai, A. K., Chauhan, B. S., Cho, H. M. and Kumar, N. (2013). A study on the performance and emission characteristics of a diesel engine fuelled with linseed oil and diesel blends. *ASME 2013 International Mechanical Engineering Congress and Exposition*.

Sathish, R., Suresh Kumar, K. and Velraj, R. (2015). Optimization of biodiesel production from Manilkara zapota (L.) seed oil using Taguchi method. *Fuel*, 140, 90-96.

Sakthivel, S., Halder, S. and Gupta, P. D. (2013). Optimisation of process variables for production of biodiesel in packed bed reactor using response surface methodology. *International Journal of Ambient Energy*, 34, 83-91.

Shahedi, M., Yousefi, M., Habibi, Z., Mohammadi, M. and Ashabi, M.A. (2019). Co-immobilization of Rhizomucor miehei lipase and Candida antarctica lipase B and optimization of biocatalytic biodiesel production from palm oil using response surface methodology. *Renewable Energy*, 141, 847-857.

Sharma, A., Singh, Y., Singh, G. K., Habte, A. T. and Singh, N. (2019). Production of polanga methyl esters and optimization of diesel engine parameters through response surface methodology approach. *Process Safety and Environmental Protection*, 121, 94-102.

Sharma, A., Muqeem, M., Sherwani, A. F. and Ahmad, M. (2018). Optimization of diesel engine input parameters running on Polanga biodiesel to improve performance and exhaust emission using MOORA technique with standard deviation. *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, 40**,** 2753-2770.

Singh, Y., Sharma, A., Singh, G. K., Singla, A. and Singh, N. (2018). Optimization of performance and emission parameters of direct injection diesel engine fuelled with pongamia methyl esters-response surface methodology approach. *Industrial Crops and Products*, 126, 218-226.

Singh, Y., Farooq, A., Raza, A., Mahmood, M. A. and Jain, S. (2017). Sustainability of a non-edible vegetable oil based bio-lubricant for automotive applications: A review. *Process Safety and Environmental Protection*, 111, 701-713.

Singh, Y. (2015). Tribological behavior as lubricant additive and physiochemical characterization of Jatropha oil blends. *Friction*, 3, 320-332.

Singh, Y., Sharma, A., Tiwari, S. and Singla, A. (2019). Optimization of diesel engine performance and emission parameters employing cassia tora methyl esters-response surface methodology approach. *Energy*, 168, 909-918.

Singh, P., Goel, V. and Chauhan, S. R. (2017). Impact of dual biofuel approach on engine oil dilution in CI engines. *Fuel*, 207, 680-689.

Singh, R., Chaudhary, R., Pandey, R., Maji, S., Babbar, A. and Chauhan, B. (2012). Performance evaluation of an air cooled diesel engine fuelled with neat neem oil and diesel blends. *Journal of Biofuels*, 3, 58-64.

Tacias-Pascacio, V. G., Torrestiana-Sánchez, B., Dal Magro, L., Virgen-Ortíz, J. J., Suárez-Ruíz, F. J. and Rodrigues, R. C. (2019). Comparison of acid, basic and enzymatic catalysis on the production of biodiesel after RSM optimization. *Renewable Energy*, 135, 1-9.

Yatish, K. V., Lalithamba, H. S., Suresh, R. and Harsha Hebbar, H. R. (2018). Optimization of bauhinia variegata biodiesel production and its performance, combustion and emission study on diesel engine. *Renewable Energy*, 122, 561-575.

Digital Super Divide: Humanity on the Crossroad

K. K. Minocha, FIETE
Anusandhan Apartment, Plot 22, Sector 6, Dwarka, New Delhi 110075 India
kkminocha@gmail.com

Abstract -- **In fast expanding digital era accelerated by frontier technologies like AI, Block chain, AR/VR, Robotics etc raise high hopes and hype. Perhaps only connected half the world populations can benefit and other unconnected half is denied. Proponents of digital ICTs and broadband tirelessly continue to speak and focus only on socio economic benefits for nations, nudging the governments to drive forward digital agenda. It is time to pause and reflect. The use of smart technology is a critical part of our evolution, but it needs to achieve mass relevance, as 30 years of global internet experience reveals. Nations find themselves on the cross roads of digital highways. Various kinds of divides always existed ever since humanity.**

But super divide - DIGITAL DIVIDE is much more frightening as it is increasing everyday very fast at exponential rate due to fast expanding technologies with rate of growth being far too greater than rate of technological diffusion and adoption amongst masses. Damaging power of digital divide is becoming too forceful to ignore. Simultaneously, owing to excessive digital consumption several damaging unintended consequences have arisen. Digital era has created new complexities/ dimensions for both heavy consumers and non consumers which merit serious attention and holistic look.

Article proposes two urgent studies to focus on damaging effects and assessment of digital era while proponents can continue focus on benefits of ICTs.

Keywords: Digital divide, ICT, World wide web, Digital transformation, Teledensity, Sustainable development goals, Urban-rural divide, PURA, E-Indexes, Digital Evolution index, Digital taxation, Usage asymmetry, Digital citizen, Digital wisdom

I. INTRODUCTION

TODAY most aspects of our lives are technology-oriented. In the digital age, information is always at our finger tips. Every day we communicate with mobile devices, use GPS to get directions, and socially interact on Face book and Twitter. The way we make decisions is getting more sophisticated because of digital technologies. Truth is about half the world's people access and use the Internet. The other half do not.

Plethora of digital jargons-Offshoots of digital era

One often hears a plethora of jargons in world of Digital technologies *viz*. digital life, digital communication, digital cities, digital government, digital divide, digital identity, digital signatures, digital libraries, digital money, digital companies, digital marketing, digital space, digital detox, digital death, digital erosion, digital fasting, digital habits, digital law, digital privacy, digital culture, digital skills, digital fluency digital foot prints, digital university, digital education, digital citizenship, digital games, digital fakes, digital natives, digital immigrants, digital citizens, digital Etiquette or Netiquette, digital Rights and Responsibilities, digital transformation, digital empowerment, digital Health and Wellness, digital Security and Protection digital planet, digital wisdom, etc. constitute various aspects of digital ecosystem and each jargon carries deep message issues for society.

The list seems ever growing as new technologies frequently evolve in digital space creating a expanding digital universe.

This digital ecosystem impacts individuals citizens, public, institutions, business and governments as users as well as non users as stakeholders.

Frontier technologies like AI, Block chain, AR/VR, Robotics etc raise high hopes and hype which perhaps only connected half the world populations can benefit and other unconnected half is denied. Digital universe is uneven.

The article discusses some of above aspects of Digital era along with contradictions digital phenomena brings in with Good, Bad & Ugly aspects and its unintended consequences.

The Digital Era: 30 years after the invention of the internet, only half of the world is connected. Internet is open conduit, it is capable of transmitting anything that can be put into digital form. It matters because of its three important characteristics *viz*. Its Global span or reach, its ability to converge everything & ability to drive innovation. By the end of 2018, half of the world's population was connected to the internet - roughly 30 years after the invention of the World Wide Web, other half often struggling to secure access.

Is this a cause for celebration or introspection? Thought contradiction of Digital modernity! Half the planet largely at bottom of pyramid who struggle with basic needs for survival (Maslow's need Hierarchy: Food, clothes and shelter) are unconnected. The use of smart technology is a critical part of our evolution, but it needs to achieve mass relevance.

Above asymmetry is cause for concern. The quotes in following paragraphs capture the essence of concern:

Observations of WWW inventor on March 12, 2019 when Internet turned 30 years old . To mark 30 years of Internet World Wide Web (www) service, its inventor Tim Berners Lee, delivered a key note speech at an event "Web@30"at the CERN said "The web is not the web we wanted in every respect."

He observed: "Web opened the way to a technological revolution that has transformed the way people buy goods, share ideas, get information and much more. It's also become a place where tech titans scoop up personal data, rival governments spy and seek to scuttle elections, and hate speech and vitriol have thrived -- taking the Web far from its roots as a space for progress-oriented minds to collaborate".

Other authors and countries based on experience too have amplified these concerns as below during early stages of internet growth:

"The empowerment of global citizenry through Information and Communication Technologies (ICTs) has drawn ample attention among development planners since the 1990s. Proponents of rural and community empowerment are encouraging governments in the developing world to take ICT initiatives, suggesting that the investment in ICTs will enable those countries to bypass more gradual development processes in order to engage their citizens on par with the developed world (Dey *et al*, 2010; Harris, 2016; Takio, 2013). But this notion is flawed for three key reasons. Firstly, it over emphasises the role of technology through simplifying the social, political and economic parameters of development. Secondly, it overlooks the needs, abilities and cultural context within which those technologies were originally introduced and interrelated; third: the importance of the political, social and economic context within which those technologies are now being applied (Mansell, 2014; Thomas, 2012)."

Besides global concerns on empowerment of citizens by ICTs similar concerns have arisen about small business as is clear from experience of OECD countries: "Information and communication technology (ICT) connectivity (PCs and Internet) is very widespread in businesses of all sizes. As is the case with all technologies, small businesses are slower than large ones to adopt new ICTs. Potential small business benefits and firm and sector-specific strategies drive the adoption and use of ICTs. Furthermore, sectors are increasingly global and dominated by large firms and the structure of their values chains and operations shape opportunities for small and medium size enterprises (SMEs). Principal reasons for non-adoption are lack of applicability and little incentive to change business models when returns are unclear.

SMEs also face generic barriers to adoption including trust and transaction security and IPR concerns, and challenges in areas of management skills, technological capabilities, productivity and competitiveness. The issues for governments are to foster appropriate business environments for e-business and ICT uptake (*e.g.* to diffuse broadband, enhance competition), and target programmes to overcome market failures to the extent that they are needed in particular areas (*e.g.* skill formation, specialised information). Governments have a range of SME e-business and Internet use programmes. However commercial considerations and potential returns are the principal drivers of small business adoption and profitable use." (OECD Report ICT, E-Business and Small and Medium Enterprises 2004). Above clearly shows thought Contradiction of Digital Modernity.

The Digital Transformation is a Double-edge sword-Good and bad Impacts. Digital transformation impacts society at several levels. On the production side of the economy, digital transformation enables the automation of business operations, yielding operational efficiencies, such as reduction of transaction costs, with an impact on productivity.

Similarly, digital transformation provides new business opportunities, impacting employment and entrepreneurship. Regarding the delivery of public services, digital transformation enhances the provision of health and education, while improving the way citizens interact with their governments. Finally, digital transformation has an impact on human relationships and individual behaviour, facilitating social inclusion and communication. It may also be noted, however, that digital transformation could also result in potential negative effects, such as workforce disruption, the disappearance of companies, cybercrime and social anomie.

For proponents of digital technologies it is worth taking note that in order to reap benefits of digital era society has to prepare and nurture all the stake holders to create balanced digital ecosystem otherwise half baked efforts are fraught with danger.

II. THREE WAVES

Digitisation refers to the transformations triggered by the massive adoption of digital technologies that generate, process, share and transfer information. Digital transformation is not a one-time event but proceeds in waves driven by technological progress and diffusion of innovations in same way as earlier waves of technological change, such as the steam engine, railroads, telegraph and automobiles that transformed society.

The first wave: (Adoption of Telecommunications and automated data processing): It comprised of introduction and adoption of what today are considered "mature" technologies. Examples include: Management information systems aimed

at automating data processing and applied to monitoring and reporting of business performance, Telecommunications technologies such as broadband (fixed and mobile) and voice telecommunications (fixed and mobile) which enable the remote access of information.

The Second wave (Diffusion of internet): It entailed the diffusion of the Internet and its corresponding platforms (search engines, marketplaces),which enable the networking of enterprises to consumers and enterprises among themselves for purchasing of supplies, and distribution of output.

The Third wave -Advance ICTs (*Rides on first and second wave*): It comprises of adoption of wide range of advanced technologies, such as big data/analytics, Internet of Things, robotics, sensors, and artificial intelligence, and is aimed at enhancing information processing and the quality of decision making, while further automating routine tasks within business enterprises and governments. These technologies are not typically adopted in a stand-alone fashion but are integrated with the mature technologies characteristic of the first and second waves. Diagram below captures these waves.

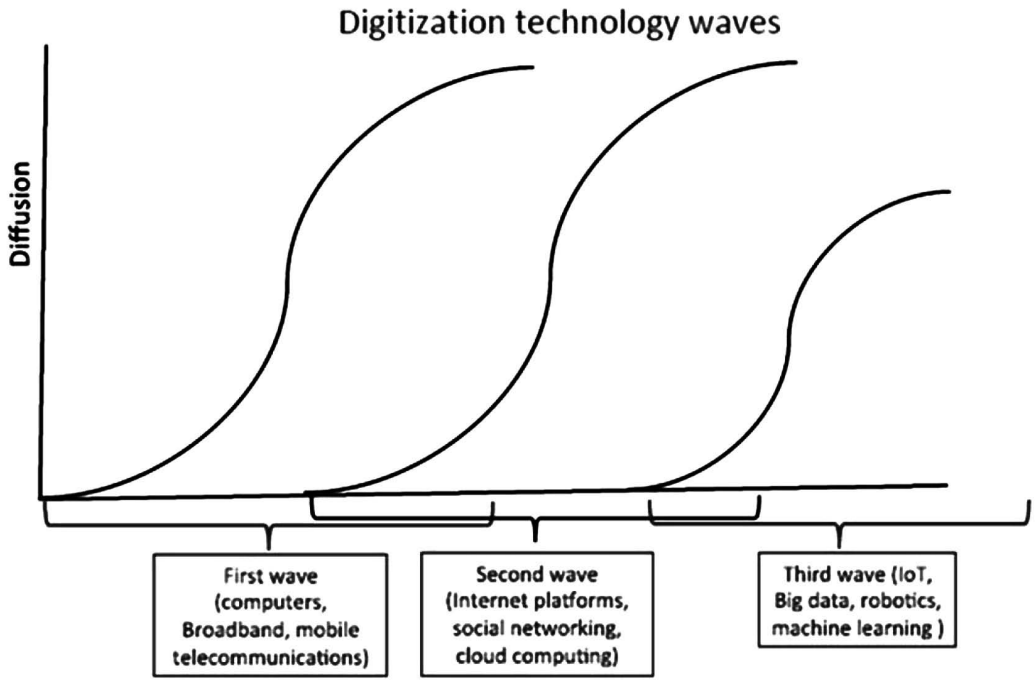
Today AI holds biggest potential to benefit society and simultaneously multiply (fast forward) prevailing digital divide.

III. RELEVANCE OF BROADBAND PENETRATION TO ECONOMIC GROWTH

Broadband is a key driver of economic growth and the competitiveness of nations. Broadband is a General Purpose Technology having a major impact on the way in which we live and work. As per often quoted World Bank research, *viz*. the '*Information and Communication for Development 2009'* suggests that the impact of broadband to economic growth is indeed substantial, much more profound than comparable narrowband or voice-based ICTs, providing a boost of 1.38 % points on GDP growth in developing countries for every 10% points increase in broadband penetration.

*Global importance attributed to power of ICTs and Digital technologies in achieving MDGs and SDGs:*In the year 2000 serious use of ICTs has been emphasised by United Nations to meet Eight MDGs (Millennium Development Goals) to be achieved by 2015. MDGs, approved by UN represent the first global development effort to provide a better future for the millions of poor citizens, with clear goals and a time frame. Four of the MDG goals pertain to health and nutrition; they are: eradication of extreme poverty and hunger; reducing child mortality, improving maternal health, and combating HIV/ AIDS, malaria and tuberculosis. Use of ICTs are part of MDG and have an impact on other MDGs.

Target 18 of MDG goal 8 mentions the following: In cooperation with the private sector, make available the benefits of new



Source: SOCIAL AND ECONOMIC IMPACT of digital transformation on the Economy - ITU

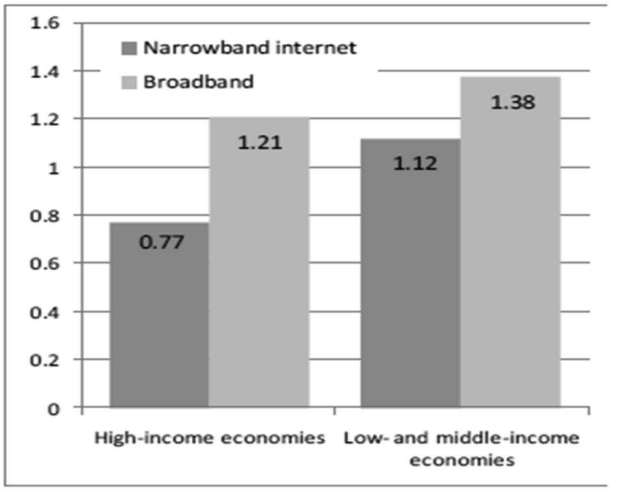


Figure 1: Impact of a ten per cent increase in penetration of selected ICTs on GDP per capita
Source: Adapted from World Bank (2009) "*Information and Communication for Development: Extending Reach and Increasing Impact*".
Note: Based on an analysis of 120 economies, 1980-2006.

Source: World Bank

technologies, especially information and communications technologies.

Table 1 below shows some examples as to how ICT can assist in achieving other MDGs.

TABLE 1 -- HOW ICTs CAN HELP ACHIEVE THE MDGs

Goal/Target	**Role of ICTs**
1. Eradicate extreme poverty and hunger	Increase access to market information and reduce transaction costs for poor farmers and traders
2. Achieve universal primary education	Increase supply of trained teachers through ICT- enhanced distance training
3. Promote gender equality and empower women	Deliver educational and literacy programmes specifically targeted to poor girls and women using appropriate technologies
4.Reduce child mortality 5. Improve maternal health 6. Combat HIV/AIDS, malaria and other diseases.	Increase access of rural care-givers to specialist support and remote diagnosis. Enhance delivery of basic and in-service training for health workers. Increase monitoring and information sharing on disease and famine
7. Ensure environmental sustainability	Remote sensing technologies and communications networks permit more effective monitoring, resource management, mitigation of environmental risks.

Source: TRAI National Broadband plan 2010

To quote from Millennium Development Goals report, 2010 (UN MDG website) "a challenge in bringing more people online in developing countries is the limited availability of broadband networks. Many of the most effective development applications of ICT, such as telemedicine, e-commerce, e-banking and e-government, are only available through a high-speed Internet connection. But a significant divide exists between those who enjoy fast access to an online world increasingly rich in multimedia content and those still struggling with slow, shared dial-up links."

In India we have achieved mixed success in meeting MDGs coupled with lack of broadband.

UN's Sustainable Development Goals (SDGs)
Encouraged by the progress with MDGs, UN General Assembly approved the more ambitious SDGs to be achieved by 2030. The progress under the MDGs, the way forward and prospects of achieving SDGs, globally and in India is reviewed by UN.

Under SDG no.9 "Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation" and

SDG target 9.c states "Significantly increase access to information and communications technology and strive to provide universal and affordable access to the Internet in least developed countries by 2020".

To achieve SDGs nations must ensure availability of advanced ICTs backed by reliable BB connectivity to every citizen with sense of urgency given fast advances in ICTs.

Various kinds of Divides- Digital divide is not a cause but it is a symptom of other divides. Divides always existed for long ever since existence of humanity. Now it has got conditioned to various types of divides/discriminations *viz*. Gender divide, Racial divide, Economic divide, Geographic divide, Rural urban divide, Information divide, Knowledge divide, Accessibility divide etc means there are significant gaps between 'Haves' and 'Have nots'. In the Digital age latest avatar of divides has been called DIGITAL DIVIDE**.**

The "digital divide"(DD) refers to the fact that certain parts of the population have substantially *better opportunities to benefit from the new DIGITAL economy* through intensive use of ICTs than other parts of the population which do not have access to ICTs putting them to position of disadvantage and inequality.

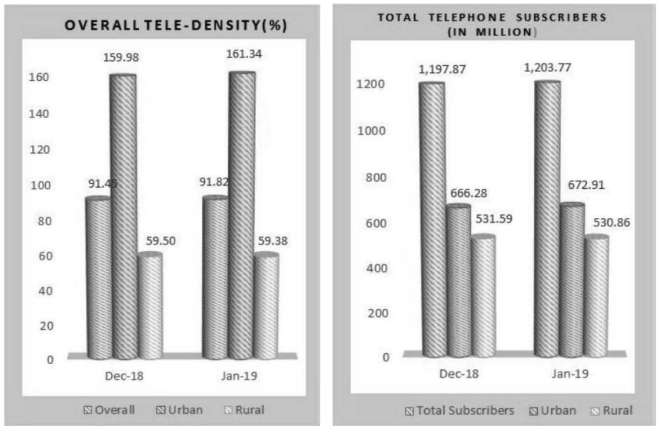
DD is any uneven distribution in the access to, use of, or impact of information and communication technologies (ICT) between any number of distinct groups; these groups may be defined based on social, geographical, or geopolitical criteria, or otherwise.

The *divide within* countries may refer to inequalities between individuals, households, businesses, or geographic areas, usually at different socioeconomic levels or other demographic categories. The divide between differing countries or regions of the world is referred to as the global digital divide, examining this technological gap between developing and developed countries on an international scale.

DD is *not a cause* but it is a *symptom* of other divides. It is kind of SUPER DIVIDE or Divide of all Divides much more powerful divide than other divides.

III. TRAI INDICATORS AND URBAN-RURAL DIGITAL DIVIDE

Latest statistics on Teledensity Rural and urban and Broadband (BB) speeds 512 Kbps or higher subscriptions released by TRAI are shown in Figures below:



Source TRAI Teledensity indicator Press Release No. 22/2019 dt 20 March 2019

Source: TRAI Broadband (connections down load speed >=512 Kbps

Percentage Teledensity growth during past decade is shown in table below. The urban areas enjoy more than 100% density *i.e.* more than one phone per person for since long time.

Year	Urban teledensity %	Rural teledensity%	Overall teledensity %
2010	122	25	53
2018	161	59	92

Overall Internet penetration due to growing no. of connections may be rising, but the urban-rural digital divide remains a reality in India. Today 512 Kbps is snail speed and no longer eligible to be called BB.

India lives in villages -Digital Divide or Urban rural divide *An unfinished agenda* Though 69 % of Indian population lives in rural areas; reliable high speed broadband facility is limited to metro and major cities. The country's overall teledensity has crossed 90%, however, its irony – more than 40,000 villages still do not have access to basic mobile services *viz*. 2G mobile coverage even today. This remains unfinished agenda for even universal mobile coverage of villages leave aside broadband coverage. Availability of broadband is critical for development of rural areas.

Rural India always takes a backseat. To date, most development theory and practice have focused on either "urban" "rural" issue with little consideration of the fact that needs context, and Ecosystem for rural are different with some areas of intersection between the two.

IV. PURA - *ASPIRATIONS!*

Provision of Urban Amenities to Rural Areas (PURA) is a strategy for rural development in India. This concept was given by former president Dr. A.P.J. Abdul Kalam. To make villages self-sufficient. PURA proposes that urban like infrastructure and services be provided in rural hubs to create economic opportunities outside cities. Physical connectivity by providing roads, electronic connectivity by providing communication networks, and knowledge connectivity by establishing professional and technical institutions will have to be done in an integrated way so that economic connectivity will emanate. PURA programs in several states since 2004, largely remains work in progress leaves much to be desired.

Description of Urban scene - Scene of Digital Abundance-Hyper connected -Over served-Over consumption (Multiple screens-Multiple SIMS per user). Many of us living urban areas enjoy variety of digital services through Broadband internet connectivity so many ways and choices. These include both Wired OFC (FTTH),Cable TV (Fixed) and Wireless 2G, 3G and 4G (Mobility) along with free Wi Fi connectivity at airports, railway stations, smart cities using variety of devices smart phones, smart watches, tablets, lap tops, TV etc and take it for granted. Though we have a choice of multiple connectivity providers TELCOs and Digital Service providers(SPs) called Over The Top services(OTTs).

Despite fierce competition amongst TELCOs some deficiencies in QoS (*e.g.* call drops, poor coverage) prevail in urban areas too. To compensate for such deficiencies we are able to avail services of more than One TELCO because of highly competitive low tariffs using 2 to 3 phones per individual (dual SIM phones) at very prices cheaper than food. Many a times we are also pampered with free services packages for months together.

Frontier technologies like Artificial intelligence, Block chain, cloud technologies etc puts urban always first in advantageous

position. Further we eagerly await much talked about 5G which will provide internet of things (IoT) such as sensors and machines on large scale etc (Industrial Revolution 4.0).

One may also say urban is land of "Haves in abundance". Even brief spells of Digital services implies Pain for digitally conditioned urban (Of course Long outages too). This online abundance is available round the clock (24x7). Many of us heavily consume digital services so much that we feel suffocated due to temporary lack/outage of it even for few minutes be it due to No coverage (Not spots) in running train or basements etc. or when Internet slows down in heavily crowded areas such as malls, cinemas etc.

Pain of long outages of Facebook FB and Instagram felt by users world over: On 13 March 2019 a hash tag started trending on Twitter: #Facebook down. FB a very popular social media site and its sister, Instagram, suffered a serious outage and restored after about 2 days time. Some users weren't able to log into their accounts at all while others were experiencing limited functionality. Impact was wide scale global in nature.

It was the worst disruption to the platform since 2008 when FB user numbers were 150 million - compared to 2.3 billion current users on the social network. Outages for large internet platforms like FB, You Tube are not unusual, but in the past they have typically restored service within an hour or two. During and after the outage, speculation was rife that FB experienced a cyber-attack *viz*. A distributed denial of service (DDoS) attack, where a website is taken offline because an attacker is flooding it with traffic. However FB in response denied this.

FB also earned bad name in year 2018 that has seen it be a victim of several successful hacks and data leaks. Because of heavy dependence Internet outages like this and also Internet shutdowns due to law and order reasons are frowned upon seriously by both individual users and business user community who have impact on their revenues and including FB platform which earns through advertisement.

Several initiatives in action are Digital India, Make in India, Start up India, Skill India, coupled with Aadhar, JAM Trinity, UPI payment platforms also the *Policy incentives* from several State Governments for ICTs urban digital consumption is rising very fast with their daily life depending on internet 24x7 including while on the move. But it has largely remained urban phenomena with little of Rural as BharatNet (Pillar no.1 of Digital India programme) is still work in progress and facing its own operational challenges on rural turf.

Description of Rural scene: Under connected, or poorly connected underserved and some Unconnected, On the contrary these areas lack above facilities because of sparsely located population or remote and difficult geographical areas or poor affordability of rural areas resulting provision of services uneconomical leading to lack of business case for TSPs. This leads to migration of people to urban areas.

Statistics capture the divide w.r.t. to number of telephone connections and for BB connections position is much worse.

Three Aspects of digital inclusion and digital divide - Digital divide is increasing rampantly everyday-estimating its damaging power-Need for reverse study-Two studies proposed here.

Divide is most researched topic amongst Economists scholars etc. There are 3 aspects of digital inclusion: (1) Access: Availability and Affordability (2) Adoption: Digital literacy (3) Application: Workforce development, Education, Healthcare, Civic engagement.

If above are not addressed they lead to Digital exclusion called divide in society. There are 3 types divides *viz*. (1) Economic Divide (2) Usability Divide (3) Empowerment divide are discussed below:

Economic divide: It is obvious and simplest form of DD is manifested in the fact that some people can't afford to buy a computer or user devices and connectivity. Although world over politicians always talk about this point and use it as tool to attract votes. In developed countries, it is reducing fast but underdeveloped countries will continue to face a challenge including our own for some more time to come. Many commentators view this simplistic way in purely economic terms *i.e.* the first aspect. As per them give people free connectivity and laptops/phones and rest will fall in line. This is wrong assumption. Ensuring technology adoption is an essential condition to achieve outcomes.

However remaining two other types of divide will have much greater impact in the years to come.

Usability Divide: Matter does not end by merely providing connectivity and user devices. Far worse than the economic divide is the fact that even today technology remains so complicated for many people are not able to use a computers and smart devices even when given free. Many others can use computers, but don't achieve the modern world's full benefits because most of the available services are too difficult for them to understand. Cyber security frauds concerns and experiences deter them to use critical online services like banking payment systems etc. Usability divide exists both in developed and underdeveloped countries.

Almost 40% of the population has lower literacy skills, and yet few websites follow the guidelines for writing for low-literacy users. Even government sites that target poorer citizens

are usually written at a level that requires a college degree to comprehend. Lower literacy is the Web's biggest accessibility problem, but hardly any attention is paid to this massive user group. Senior citizens face the second-biggest accessibility problem, but again there is little interest in the guidelines for making websites easier for old users.

V. USAGE ASYMMETRY

Though the economic divide is closing in many parts, perhaps there is little progress on the usability divide. Usability is improving for high-end users. For such user segment, websites and applications like e-commerce, social platforms, payment platforms, Taxi hailing get easier every year, generating vast profits for website owners and OTT providers. That's all great news for high-end users, but the less-skilled 70% or more of users have seen little in the way of usability improvement. We know how to help these users -- but simply not doing it.

Empowerment Divide: Society has knowledge needed to bridge the usability divide, and one hopes in due course that this aspect is closed. The empowerment divide, however, is the hard one: even if computers and the Internet were extraordinarily easy to use, not everybody would make full use of the opportunities that such technology affords.

Participation inequality is one aspect of the empowerment divide that has held constant throughout all the 30 years of Internet growth: in social networks and community systems, about 90% of users don't contribute, 9% contribute sporadically, and a tiny minority of 1% accounts for most active contributions.

Example below illustrate Empowerment divide -Many users don't know even how to use internet search engines.

Research shows that many users don't know how to use search to truly master the Web. People don't understand advanced search features, they rarely employ query reformulation, and many uncritically select the first search results.

Similarly many people are not aware large number of features available in smart phones which could enhance their productivity.

Relative position of progress amongst above said 3-Divides: In short ultimately, one can be hopeful about the 'economic divide', which is closing rapidly in many countries. However the 'usability divide' will take longer to close, but at least we know how to handle it -- it's simply a matter of will to do so. But about the 'empowerment divide', one gets pessimistic may expect that it will only grow more severe in the future.

Today In developed world the economic divide is a non-issue, but the usability and empowerment divides alienate huge population groups who miss out on the Internet's potential, but in developing world all the three pose a serious challenge.

VI. REASONS WHY DIGITAL DIVIDE EXISTS?

While there is no fixed answer to the question "Why digital divide exits?" answer of why digital divide exists is complex, as it depends on several factors, and many of which cuts across the globe. On the other hand, some issues are very specific to individual country, region or state. Major factors that cause digital divide includes:

Fast changing Technologies: The internet is expanding very quickly, and not all countries--especially developing countries- are able to keep up with the constant changes. Most of the developing countries including India have low Internet penetration, leading to denial of access to many web based information and latest technology.

Slow diffusion of new technologies: Another reason of digital divide is economic inequality. The digital divide comes from slow diffusion of new technologies to selective sections in society or countries. Primarily, wealthier peoples and countries carry out these experiments with new technologies because they have disposable income. But over time, these divides closes considerably as the technology becomes less expensive and more tested.

Socio-Economic reasons: Even if there is enough diffusion of technology and access to information and knowledge, not all people can afford to reach the knowledge hub. For example, not all kids are getting as much technical education as others because lower socio-economic areas cannot afford to provide schools with bundles of computers. For this reason, kids are being separated and not receiving the same chance as others to be as successful.

In Multilingual India unique case Language divide challenge: One of the important criteria of digital divide is language barrier. Most of the contents in the web is either written in English or other foreign language, which Indian are not well versed. As a result getting benefit out of these contents are not possible. Diverse country like India with large number of regional languages faces challenge this divide no other country has such intense challenge.

Advanced ICTs widen the divide. Recent phenomena of Big data, Cloud computing, IoT Industry 4.0, Block chain AI (Artificial Intelligence) technologies etc constitute advanced ICTs . AI is increasingly becoming the veiled decision-maker of our times.

Investments in infrastructure for advance ICTs: Harnessing the benefits of advanced ICTs requires appropriate infrastructures, services and skills. Networks will have to support diverse

quality-of-service demands from applications and users while delivering robust and ubiquitous connectivity. This requires the roll-out of wireless Internet of Things platforms, and relying on network virtualization and improved fiber connectivity. Moreover, it will require the development of advanced ICT skills among users.

Advanced ICTs raise important concerns over next-generation digital divides. Network operators and users will have to adapt their business models to take advantage of the opportunities of the digital transformation. Thus, policy makers and regulators are called upon to create conditions facilitating entrepreneurial experiments and innovation.

Thus Digital divide is double edge sword - a Complex and dynamic Phenomenon. To close the divide, society must respond at very fast rates with matching preparations and local solutions as preparedness for fire fighting is needed.

VII. TWO URGENT STUDIES PROPOSED TO ADDRESS IMBALANCE IN DIGITAL ECOSYSTEM

Proponents of Digital ICTs and Broadband always speak and focus on socio economic benefits for nations and nudge the governments to drive forward digital agenda. Now with 30 years of global experience with internet nations find themselves on the cross roads of digital highways. Now time has come to reflect and introspect.

Today in era of fast expanding technologies where rate of technological growth is far too greater than rate of technological diffusion and adoption Digital divide is increasing everyday very fast exponentially. Damaging power of divide is becoming too forceful to ignore. Digital divide is increasing rampantly everyday due to fast proliferation of advanced ICTs in uneven pockets causing immense damage to have nots and also to have access who are not able to harness it fully (suboptimal use) estimating its damaging power-Need for studying the unintended consequences that arise from this.

Therefore it is proposed that there is now a urgent need to carry out a reverse study to assess extent of socioeconomic damage caused by rampantly growing divide by reputed agencies like World Bank, ITU and each of affected nations themselves to localise context similar to world bank's previous study of assessing economic benefit resulting from increase in broad penetration often quoted by experts in seminars (See above Relevance of Broadband penetration to Economic growth source world bank).

Proposed Reverse study should exclusively focus on socio economic damaging impact of digital divide which could read something like this to rhyme with prevailing one : A 10 percent increase in the digital divide rate in developing countries is associated with a X percent damage in GDP per capita and damage is far too higher than the equivalent relationship for developed countries at Y %.

It's no brainer that damage value X will be far greater than it's equivalent Y .The intent is to estimate values of X and Y and then quote and sing this song in every seminar to lay emphasis on damage side so that resultant damage caused by digital divide is brought under check, duly contained and in control of instead of continuing with rampant accumulation similar to irreversible damage caused to earth by plastic, fertilizers, nuclear technologies etc.

Henceforth it will also help in calibration and keep in check induction of digital ICTs by champions and proponents is backed with simultaneous adequate preparation proactively for end to end net positive impact on the lives of users and nations and not reckless growth and reactive response.

It will go a long way in awakening consciousness of proponents of digital technologies, governments and nations who have started buying the idea that digital technologies are panacea. There is also need for commissioning second study to examine and measure the damaging effect caused by excessive digital consumption bordering abuse or addiction in all walks of life. Objectives and parameters of study can be suitably framed for meaningful outcomes. The idea of these two proposed studies to bring balance to digital ecosystem for intended welfare and avoid unintended consequences for nations and the planet.

VIII. DIGITAL NATIVES, DIGITAL IMMIGRANT, DIGITAL FLUENCY TO DIGITAL WISDOM

As Digital era evolves globally some interesting insights such as Digital Natives, Digital Immigrant, Digital Fluency to Digital wisdom have cropped up. Digital native is a term coined by an Educationist Mark Prensky in 2001. It was used to describe an individual who was born after the widespread adoption of digital technology and grew up in the digital age. Digital natives are comfortable with technology and computers at an early age and consider technology to be an integral and necessary part of their lives. For North America, most people born prior to 1980 are considered digital immigrants. This cut off year 1980 is arbitrary without any rationale.

Teenagers and children today are generally considered to be digital natives as they mainly communicate and learn via Internet, computers and mobile devices computers, mobiles, texting. This exposure to technology in the early years is believed to give digital natives a greater familiarity with and understanding of technology than people who were born before it was widespread.

A digital immigrant is an individual who was born before the widespread adoption of digital technology. The term digital immigrant may also apply to individuals who were born after

the spread of digital technology and who were not exposed to it at an early age. Not all children born today are digital natives by default.

Digital immigrants are the opposite of digital natives, who have been interacting with technology since childhood. Digital immigrants are believed to be less quick to pick up new technologies than digital natives. However, this classification of people into Natives and immigrants is controversial.

IX. MEASURING THE INFORMATION SOCIETY IN DIGITAL ERA

It is often said what ever gets measured gets done. Measuring the Information society presents its own challenges in ever evolving fast paced Digital world. e-readiness (electronic readiness) index is a measure of the degree to which a country, nation or economy may be ready, willing or prepared to obtain benefits which arise from information and communication technologies (ICTs).

An index is comprised of multiple relevant indicators/ variables that can be concluded into a single value to evaluate effectiveness on a certain subject. For this purpose, composite e-indexes have been developed to measure the levels of ICT Adoption in different countries all over the world by many globally recognized organizations and institutes have introduced a number of indices related to telecommunications & ICT sector (See list below).

e-readiness is considered one of the main faces of development for any country and represents the transformation of society including the movement from traditional methods to more modern ways of thinking or dealing with health, education and production.

The ICT Development Index (IDI), published annually is a composite index that combines 11 indicators into one benchmark measure. It is used to monitor and compare developments in information and communication technology (ICT) between countries and over time. Every year ITU conducts country wise global study of development of ICTs in world and ranks them. From 2018 onwards index has been refined to contain 14 indicators instead of 11 to add further insights into the performance of individual countries and the relative performance of countries at different development levels. The IDI is a standard tool that governments, operators, development agencies, researchers and others can use to measure the digital divide and compare ICT performance within and across countries.

India's IDI & Digital communication Policy aspirations: Government is seized of poor status in world ranking. Therefore it has included IDI improvement as one of policy objectives in Telecom policy *i.e.* to improve its IDI ranking to Top 50 nations by 2022 from ranking of 134 in 2017.Digital communications policy 2018 states in its objectives -"Propelling India to the Top 50 Nations in the ICT Development Index of ITU from 134 in 2017".

Network Readiness Index (NRI) framework of WEF: The World Economic Forum developed Networked Readiness Index (NRI), also called Technology Readiness, developed in 2002. It measures the propensity for countries to exploit the opportunities offered by information and communications technology (ICT). NRI enables evaluation of a country's level of readiness to obtain and participate in the benefits of ICT. Objective of NRI is to assess the impact of ICT on the competitiveness of nations.

TABLE 2 -- LIST OF MOST WIDELY USED E-INDEXES

Short name	Full name	Author	First publication	Number of countries
ISI	Information Society Index	IDC	1997	53
ERI	E-Readiness Index	EIU	2000	70
KEI	Knowledge Economy Index	WEF	2005	140
EGDI	E-Government Development Index	UNPAP	2002	182
IDI	ICT Development Index	ITU	2002	154
DAI	Digital Access Index	ITU	2003	178
TAI	Technology Achievement Index	UNDP	2001	72
NRI	Networked Readiness Index	WEF	2002	148
DOI	Digital Opportunity Index	ITU	2005	181
ICT-OI	ICT Opportunity Index	ITU	2005	183
ICT-DI	ICT Diffusion Index	UNCTAD	2006	180
GII	Global Innovation Index	INSEAD	2007	143
GCI	The Global Competitiveness Index	WEF	2004	144
IS	Infostates	ORBICOM	2003	183

Source: Some Aspects of ICT Measurement: Comparative Analysis of E-Indexes, Kateryna Kononova National University, Kharkov, Ukraine

Digital Evolution Index Map- (Speed of Change in Digital world)-Mapping momentum around the world instead of tabular ranking: Digital Evolution Index captures Speed of Change in Digital world as in the digital age, speed is the new scale. An interesting framework to track digital evolution in a fast changing Digital landscape called Digital Evolution Index developed by HBR in collaboration with Fletcher School at Tufts University and Master card in 2015 to trace the emergence of a "digital planet," and to map Digital Momentum around the World. By measuring each country's current state of digital evolution and its pace of digital evolution over time (2008-2015, HBR Fletcher-Master cardteam created a chart or a map of our digital planet (see chart below).

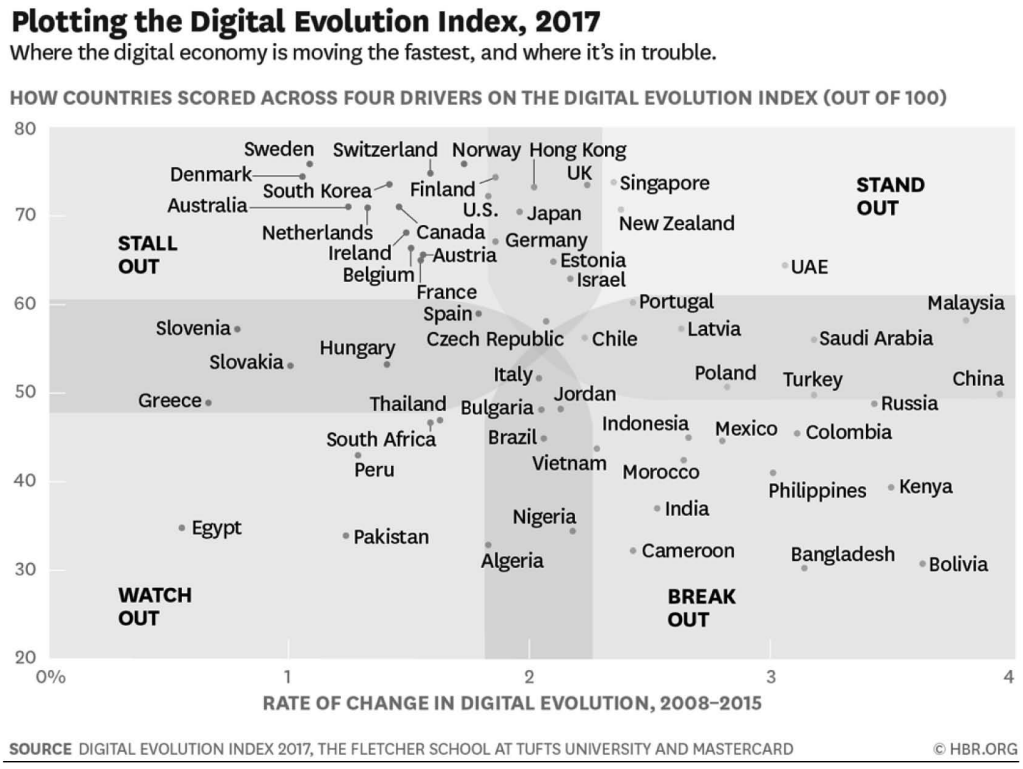
Charts Horizontal axis represents Rate of change of in digital evolution and Vertical axis How countries scored on the Digital evolution index (Out of 100).Countries on this chart fall into four zones: Stand Out, Stall Out, Break Out, Watch Out. Some countries are at the border of multiple zones.

Stand Out countries are highly digitally advanced and *exhibit high momentum*. They are leaders in driving innovation, building on their existing advantages in efficient and effective ways. To stay ahead, these countries need investments to keep their innovation engines in top gear and generate new demand, failing which they risk stalling out. Stall Out countries enjoy a high state of digital advancement while *exhibiting slowing momentum*. The five top scoring countries in Stall Out zone in the DEI 2017 ranking — Norway, Sweden, Switzerland, Denmark, and Finland, reflecting the challenges of sustaining growth. Stall Out countries may look to Stand Out countries for lessons in sustaining innovation-led growth. Countries in the Stall Out zone can put their maturity, scale, and network effects to use to reinvent themselves and grow.

Break Out countries are low-scoring in their current states of digitalization but are *evolving rapidly*. The *high momentum of Break Out countries* and their significant headroom for growth would make them highly attractive to investors. Break Out countries are often held back by relatively weak infrastructure and poor institutional quality. India occupies in this zone. Break Out countries have the potential to become the Stand Out countries of the future, with China, Malaysia, Bolivia, Kenya, and Russia leading the pack.

Watch Out countries face significant challenges with their low state of digitalization and low momentum; in some cases, these countries are moving backward in their pace of digitalization.

Inferences from DEI 2017: Clearly two of the world's most significant economies, the U.S. and Germany, are at the border of Stand Out and Stall Out, with a third, Japan, in the neighbourhood. Digitally speaking Clearly and notably, the most exciting region in the world, is Asia, with China and Malaysia as exemplars. We can expect to see plenty of investor and entrepreneurial interest in this region; it is critical that the political institutions are stable and supportive. India, with many policy-led pushes for digitalization, including a Digital India



campaign and initiatives to give a boost to digital payments, ought to pay attention to the overall low level of evolution in the country. This can act as a drag on any initiative. Broader and more systemic changes are needed to boost digital evolution in this type of environment.

IX. DIGITAL TAXATION CHALLENGE

Digital companies in digital era manifest contradictions on taxation making it challenging - They Exploit Tax loop holes- France is first country to show the way to world. This a contradiction that while on one hand new digital companies world over themselves disrupt every sector of economy, every walk of life with their new innovations and business models resulting in death of some old companies but when it comes to their own lawful taxation in accordance with new taxation models they avoid, resist and want to stick to taxation laws meant for traditional brick and mortar business models by exploiting loop holes in existing laws and continue to fight prolonged legal battles against tax authorities in courts for years together.

Traditionally, brick & mortar companies have been taxed on their profits. Calculating profit is a complex affair, affording much scope for creative accounting in digital economy dealing with digital companies. Traditional taxation is globally disrupted in digital era.

Some experts have for long suggested taxing companies on their sales, not profits, to end tax avoidance through creative accounting. However, other experts opine this will be unfair to struggling companies that lose money despite having large sales. That is why taxes are almost everywhere levied on profits, not sales. France on 11 July 2019 enacted legislation to tax digital giant companies called 'GAFA' (Acronym for digital biggies- Google, Apple, Facebook and Amazon) at rate of 3% on their revenue, not profits. Legislation has been dubbed as- 'GAFA tax'.

India in 2016 made mild beginning which levies a 6% equalisation tax instead of a straight tax on digital online advertising platforms like Google, FB etc, limited to tax on the fees that advertisers pay to these platforms. Eventually India and many other countries too need to go France way *i.e.* tax revenues though not immediately in a calibrated way.

This is essential to tackle the completely new phenomenon of digital companies that can show massive sales, but no profits, in many countries they serve.

The rationale behind devising a separate framework to tax online service providers is this: That existing system has been disrupted by the new business models of digital giants. Unlike traditional companies, they face no import duties or non-tariff barriers when they enter foreign countries: it is impractical to levy an import tax or sales tax on data. Hence, digital activities have escaped tax. This is because the digital economy is characterized by a unique system of value creation resulting from a combination of factors such as sales functions, algorithms and personal information of users.

What distinguishes technology companies from traditional businesses is user participation in creating value, which, in turn, translates into revenue. Though using consumer data to improve businesses is not exclusive to the digital economy, the unique ability of digital businesses lies in their power to analyse big data collected via constant user interaction and data mining.

Further Digital giants can show virtually no profits in the countries they serve, and show virtually all their profits in their headquarters in a low-tax countries like Ireland or Luxembourg. This can be done by vesting intellectual property rights (IPR) with the company in the tax haven, and arranging to have all other subsidiaries pay high licensing fees for this IPR, thus transferring the bulk of profits to the tax havens. This is a problem for the US too. Trump administration is livid that US companies are salting away their profits abroad and not bringing them back to the own country.

But this is also a problem for all countries finding that digital companies have massive activities in their territory that would be taxed substantially if they were traditional sales, but escape altogether because they are digital.

Matter also came up at meeting of finance ministers G 20 nations in June 2019 where they agreed to find new ways to push ahead on compiling common rules that will close loopholes used by global technology giants like GAFA to reduce their corporate tax burden. Ideally, an international agreement should have been reached on this issue long ago. But the interests of all countries are not aligned. Hence, there is pressure for unilateral action.

X. DIGITAL CITIZENS

Digital Citizenship is a concept which helps us understand what all technology users should know to use technology appropriately and responsibly. Digital citizenship can be defined as engaging in appropriate and responsible behaviour when using technology. It encompasses digital literacy, ethics, etiquette, online safety, norms, rights, culture and more. As in any society, it is expected that digital citizens act in a certain way according to accepted norms, rules, and laws. All the online users leave a trail of their digital footprints whenever you use technology online. This includes digital cameras, smart phones, iPods, tablets, laptops and computers. A Digital Footprint is the record of digital users interaction with the digital world.

A Digital Footprint is like an invisible report card that accumulates over the years and is a general reflection of our

online activity. It also reflects personality traits of online user and is visible to anyone on social platforms may affect future employment of youngsters who should be very careful in their online activity.

XI. DIGITAL WISDOM AND TRADITIONAL WISDOM

For today's young people, using technology is as fundamental as reading was for their parents and grandparents. Technology underlies and supports everything they do. Today we all face the need to adapt quickly to a very different world, one that is growing increasingly complex and changing at an ever-accelerating pace. Technology can be used for great things, but it can also be a massive distraction. Being digitally wise involves allowing machines to do what people want them to do. It may mean being careful what we do online and inculcate digital habits which enhance and empower us. Wisdom comes from different sources: individual wisdom, collective or organizational wisdom (wisdom of crowd) and machine wisdom (artificial intelligence).

Today there is need to redefining traditional Wisdom. Our "Traditional" thinking is undergoing profound change--so much so that finding the best combination of mind and technology requires new wisdom termed as "digital wisdom." Old kinds of wisdom, like memorizing a great deal of information when young that will work for the rest of your life, or seeking a job or employer you can remain with for an entire career, no longer make sense in our contemporary world. Today information goes out of date far too rapidly. Today skills must come from multiple disciplines and sources.

Not that our "old" wisdom never counts or applies—much still does. But we need to figure out where and when the traditional wisdom does and does not work, and when it doesn't, we need to put something new in its place. *Certain things* viz. *human passion, empathy, or yearning can never be outsourced to technology.* Some of our digital enhancements may bring ethical dilemmas. With our eyes wide open to the potential benefits and harms of digital technologies, let us pursue digital wisdom in order to succeed in the twenty-first century and endeavour to become a good digital citizen.

Finally quoting from book Wisdom 2.0- The New Movement Toward Purposeful Engagement in Business and in Life "Technology is not the answer. It is also not the problem. What matters instead? Awareness, Engagement, and Wisdom.

Wisdom 2.0 addresses the challenge of our age: to not only live connected to one another through technology, but to do so in ways that are beneficial, effective, and useful.

XII. CONCLUSION

Only a half-century after 30 years the invention of the internet, and half of the world is unconnected. This creates huge divide. The digital revolution is here to stay. Digital technologies can make us live efficiently provided we use them with care and caution. Our need of digital wisdom is important and urgent. It will help us behave properly in the digital world where there are no secrets and everything is open.

Digital world era has its inherent contradictions: Good Bad and Ugly aspects. All these come together as a package. While enthusiasts can be Gung ho about turning everything and humanity into DIGITAL avatar but we have to tread digital path with caution and care. Unintended consequences of Digital era have to be guarded as in past other technological revolutions *viz*. Nuclear, Plastics, Processed Foods, Fertilizers, and their reckless abuse of these have caused vast irreversible damage to humanity and environment given us enough experience and learning to deal with digital revolution. Two studies have been proposed in article with an intent to bring balance to digital ecosystem for intended welfare and avoid unintended consequences for nations and the planet.

REFERENCES

[1] https://www.thestar.com/news/world/europe/2019/03/12/at-age-30-world-wide-web-is-not-the-web-we-wanted.html As the Internet turns 30, founder says it's 'not the web we wanted'- By Jamey Keaten The Associated Press, *March 12, 2019.*

[2] P.N.Thomas, *Digital India: Understanding information, communication and social change*, New Delhi, Sage, 2012.

[3] R. Mansell. Empowerment and or/Disempowerment: The Politics of Digital Media, *Popular Communication*, vol. 12, pp. 223-236, 2014.

[4] ICTs and empowerment in asymmetrical Power relations: a case of union Information and service centre (uisc) in Bangladesh- *Mohammad Sahid Ullah* https://www.researchgate.net/publication/315666462_ICTs_and_Empowerment_in_Asymmetrical_Power_Relations_A_case_of_Union_Information_and_Service_Centre_UISC_in_Bangladesh.

[5] ICT,E-Business and Small and Medium Enterprises OECD. https://www.oecd-ilibrary.org/docserver/232556551425.pdf?expires=1563199327&id=id&accname=guest&checksum=08CDE5D0121416597C5901A1788C4980.

[6] https://intelligence.weforum.org/topics/a1Gb0000001SH21EAG?tab=publications.

[7] Social and Economic Impact of digital transformation on the Economy – ITU. https://www.itu.int/en/ITU…/Soc_Eco_impact_Digital_transformation_finalGSR.pdf.

[8] http://ecdl.org/policy-publications/digital-native-fallacy, The fallacy of the 'digital native'.

[9] https://www.techopedia.com/definition/28094/digital-native.

[10] https://www.weforum.org/agenda/2018/08/we-cant-tell-if-were-closing-the-digital-divide-without-more-data/ "We don't have enough data to draw conclusions about the digital divide".

[11] https://hbr.org/2017/07/60-countries-digital-competitiveness-indexed 60 Countries' Digital Competitiveness, Indexed JULY 12, 2017 by Bhaskar Chakravorti, Ajay Bhalla and Ravi Shankar Chaturvedi

[12] https://www.researchgate.net/publication/220136939_Digital_Divide_in_India_Measurement_Determinants_and_Policy_for_Addressing_the_Challenges_in_Bridging_the_Digital_Divide.
[13] http://ijarcsse.com/Before_August_2017/docs/papers/Volume_7/8_August2017/V7I7-01711.pdf
Digital Wisdom Matthew N.O. Sadiku1, Adebowale E. Shadare, Sarhan M. Musa.
[14] https://www.qualcomm.com/news/onq/2012/11/15/why-we-need-digital-wisdom Why We Need Digital Wisdom.
[15] https://wisdom2conference.tumblr.com/post/76584880246/what-is-wisdom-2-0.
[16] https://www.thestar.com/news/world/europe/2019/03/12/at-age-30-world-wide-web-is-not-the-web-we-wanted.html.
[17] https://www.forbes.com/sites/kateoflahertyuk/2019/03/14/was-the-facebook-outage-a-cyber-attack/#3990d2615223.
[18] https://en.unesco.org/system/files/unesco-steering_ai_for_knowledge_societies.pdf. Steering AI and Advanced ICTs for Knowledge Societies.
[19] https://www.nngroup.com/articles/digital-divide-the-three-stages/ Digital Divide: The 3 Stages by Jakob Nielsen.
[20] http://dot.gov.in/ Digital communications policy 2018 of India.
[21] https://www.virtuallibrary.info/digital-citizenship.html. Digital citizenship .
[22] Some Aspects of ICT Measurement: Comparative Analysis of E-Indexes -Kateryna Kononova1.
[23] https://economictimes.indiatimes.com/blogs/Swaminomics/tech-what-one-deserves/ Tech What One Deserves by SA Aiyar in Swaminomics *Economic Times*, July 16, 2019.



K.K. Minocha FIETE graduated from Delhi College of Engineering in the year 1976 and MBA (Finance) from IGNOU. Professionally qualified Telecom engineer with 41 years of rich, diverse experience in various fields of telecommunications.

Belongs to 1978 batch of Indian Telecom Service. Served in Department of Telecom in variety of areas of Policy, USOF, Project installation, Maintenance, planning and management of both modern long distance telecom network and local network providing various kinds of services.

During the career visited 12 countries abroad on different assignments including training. Worked as APT Expert on Digital International Satellite system to DPR Korea at Pyongyang for digitalization of their international satellite network in the year 1999.

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/337394590

Effect of Fiber Length and Content on Mechanical and Water Absorption Behavior of Coir Fiber- Epoxy Composite

Chapter · November 2019

CITATIONS
0

READS
10

2 authors:



Mohit Mittal
Delhi Technological University
16 PUBLICATIONS **16** CITATIONS
SEE PROFILE



Rajiv Chaudhary
Delhi Technological University
70 PUBLICATIONS **178** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project Lignocellulosic fibers based composite material View project

Project Corrugation process parameters optimization View project

All content following this page was uploaded by Mohit Mittal on 20 November 2019.

The user has requested enhancement of the downloaded file.

CHAPTER - 7

Effect of Fiber Length and Content on Mechanical and Water Absorption Behavior of Coir Fiber-Epoxy Composite

Mohit Mittal [a,] *, Rajiv Chaudhary [a]

[a] *Department of Mechanical, Production, and Industrial Engineering, Delhi Technological University, Delhi, 110042, India.*

* *Corresponding author. E-mail address: mohit.30mittal@gmail.com*

Abstract

The geometrical dimensions (length and diameter) and the volume content of reinforced fibers in a polymer matrix play a significant role to affect the performance of composite materials. Therefore, to elevate the inherent strength of polymers and their probability of selection, it is highly require to reinforcing the plastics with an optimum fiber length and content. In this context, sixteen samples of coir fiber-reinforced epoxy composites (CF-EP) were made by varying the fiber length (10 mm, 15 mm, 20 mm, and 25 mm) and fiber volume content (17%, 23%, 34%, and 43%). The developed composites were characterized for mechanical properties and the water absorption behavior. The tensile, flexural, and izod impact tests were performed as per ASTM standard. Moreover, the water absorption behavior and its effect on the mechanical properties of composites have been carried out. The experimental results revealed that the tensile and flexural strengths of CF-EP increases with the increase of fiber volume content up to 23%. However, the maximum impact strength was obtained with 43% fiber volume content. The optimum length of reinforcing fibers is 20 mm at which the composite material exhibits best set of mechanical properties. The equilibrium moisture content (M_{eq}) in CF-EP composite increases from 1.8% to 5.32%, as the fiber volume was increased from 17% to 43%. Further, the results also demonstrate that the tensile and flexural properties (strength and modulus) of wet composites are lower than the dry composites.

Keywords: coir fiber, mechanical properties, fiber content, fiber length, water absorption, polymer composites

1. INTRODUCTION

In current days, the polymer based composite materials are enjoying a lot of concern from industries, academia, and research groups. The increasing of awareness towards the issues of socio-economic, technological, environmental, and energy optimization resulted in the higher usages of natural fiber reinforced composites (NFRC's) in automobiles, railways, packaging, furniture, sports goods, aerospace, construction, and marine industries. NFRC's have exceptional features such as light weight, low cost, renewability, non-hazardousness, biodegradability, easy availability, high specific strength and stiffness, corrosion resistant, safe manufacturing, and excellent acoustic, thermal, and electrical insulation than the glass fiber composites [1]. A lot of research works have been done in past on the deployment of local cheap and efficient natural fibers such as jute, banana, rice straw, bamboo, and sisal, etc. as reinforcement in polymer composites [2-6]. Among various types of cellulosic fibers, the coconut husk fiber (Coir) has received great attention due to its remarkable properties such as high toughness; hardness; stiffness; resistant to heat, salt water, microbial and fungus degradation; durability; and hardwearing qualities. The coir fiber contains relatively higher lignin content than the other commonly used natural fibers which provides it long lasting, less combustible and reliable properties. D. Verma *et al.* [7] reported that the coir is largely available in India and about 52 billion of coconuts are yearly produced in world. Therefore, there is a need to develop and explore the materials based on coir fiber. G. Das *et al.* [8] studied the effect of fiber size and content on the mechanical properties of Coir-Epoxy composites and reported that the composite which consists of 15 wt% fiber at 12 mm fiber length exhibits maximum strength. G. Rajeshkumar *et al*. [9] investigated the effect of fiber length and content on the mechanical properties of Phoneix sp. fiber-epoxy composites and found that the highest tensile and impact strengths were obtained with 40% fiber volume content at 300 µm particles size. R. Shrivastava *et al*. [10] studied the mechanical properties of Coir/Glass hybrid epoxy composites and stated on that the tensile and flexural strength of composites increased with the increase of coir fiber content. S. Nallusamy *et al*. [11] investigated the mechanical properties of coir fiber-polyester composites filled with bone and sea shell powders. V. Bhagat *et al.* [12] reported that the maximum strength of Coir/Glass hybrid epoxy composite was obtained with 10 wt% loading of coir fiber with 15 mm fiber length. G. Goud *et al*. [13] studied the effect of *Roystonea regia* fiber content on the tensile strength of epoxy based composites and reported that the maximum increment in strength of composite (32.5 MPa to 38.89 MPa) was achieved after the loading of 20 wt.% fiber content. M. Mittal *et al*. [14] studied the effect of fiber content on the thermal behavior of coir-epoxy composites and reported on that the peak decomposition temperature (T_P) of the epoxy thermoset was increased from 371.96 ^{0}C to 386.03 ^{0}C after the 23% loading of coir fiber content. U. Devi *et al*. [15] reported that the optimum mechanical performance of PALF-Polyester composites was obtained with 30% PALF content at 30 mm fiber length. M.C.N Yemele *et al.* [16] reported the increment in tensile and flexural strengths of HDPE matrix by the addition of bark fiber. N. Venkateshwaran *et al.* [17] investigated the effect of fiber length and content on the mechanical properties of banana fiber-epoxy composite and reported that the optimum weight ratio and length are 16% and 15 mm respectively. A. Sbiai *et al.* [18] studied the effect of fiber size on dynamic mechanical properties of date palm fiber-epoxy composite. V.A Alvarez *et al.* [19] studied the effect of sisal fiber loading on equilibrium moisture content of starch

based composite materials and reported that the equilibrium moisture content was increased with the increase of fiber volume fraction. El-Shekeil *et al*. [20] investigated the effect of kenaf fiber content on tensile strength of Kenaf/PU composites and concluded that the maximum value of tensile strength was obtained with 30% fiber content. B. Bakri *et al*. [21] studied the effect of coir fiber content on mechanical properties of coir/agave hybrid epoxy composites and reported on that the composite of 15 vol.% coir fiber yields higher tensile and flexural properties than the other composite samples (fiber:matrix-10:20; 20:10). To this end, the primary aim of this work is to study the effect of fiber size and content on the mechanical properties of coir fiber-reinforced epoxy (CF-EP) composites. Moreover, this work also deals with the water absorption behavior and its effect on the mechanical properties of CF-EP composites.

2. MATERIALS AND METHODS

A. Polymer Matrix Resin

A medium viscosity epoxy resin (DGEBA, LY556) with hardener (TETA, HY951) were procured from M/s Sakshi Dies and Chemicals, Delhi (India) and used as a matrix material. The important properties of the epoxy resin are mentioned in Table 1. The primary reasons to select the epoxy thermoset as a matrix material are its superior mechanical, chemical, and corrosion resistant properties.

Table 1 Properties of epoxy resin

The appearance of liquid resin	A clear pale yellow
Density (Kg/m^3)	1200-1250
Viscosity at 25^0C (cps)	550
Tensile strength (MPa)	11.14
Tensile modulus (GPa)	0.48
Flexural strength (MPa)	25.29
Flexural modulus (GPa)	1.53
Max. Elongation (%)	4.2-5.6
Impact strength (kJ/m^2)	1.2
Flashpoint	> 200^0C

(Authors observations)

B. Reinforcing Fiber

The coir fibers of density 1.24 g/cm^3 were procured from M/s Go Green Products, Chennai (India). The fibers were of long form and cut into the different lengths (10, 15, 20, and 25 mm). The maximum moisture content in coir is 7-8% and the approximate chemical composition is: cellulose (32-43%), hemicellulose (0.15-0.25%), lignin (40-45%), and pectin (3-4%). The SEM micrograph of coir fiber is shown in Figure 1.

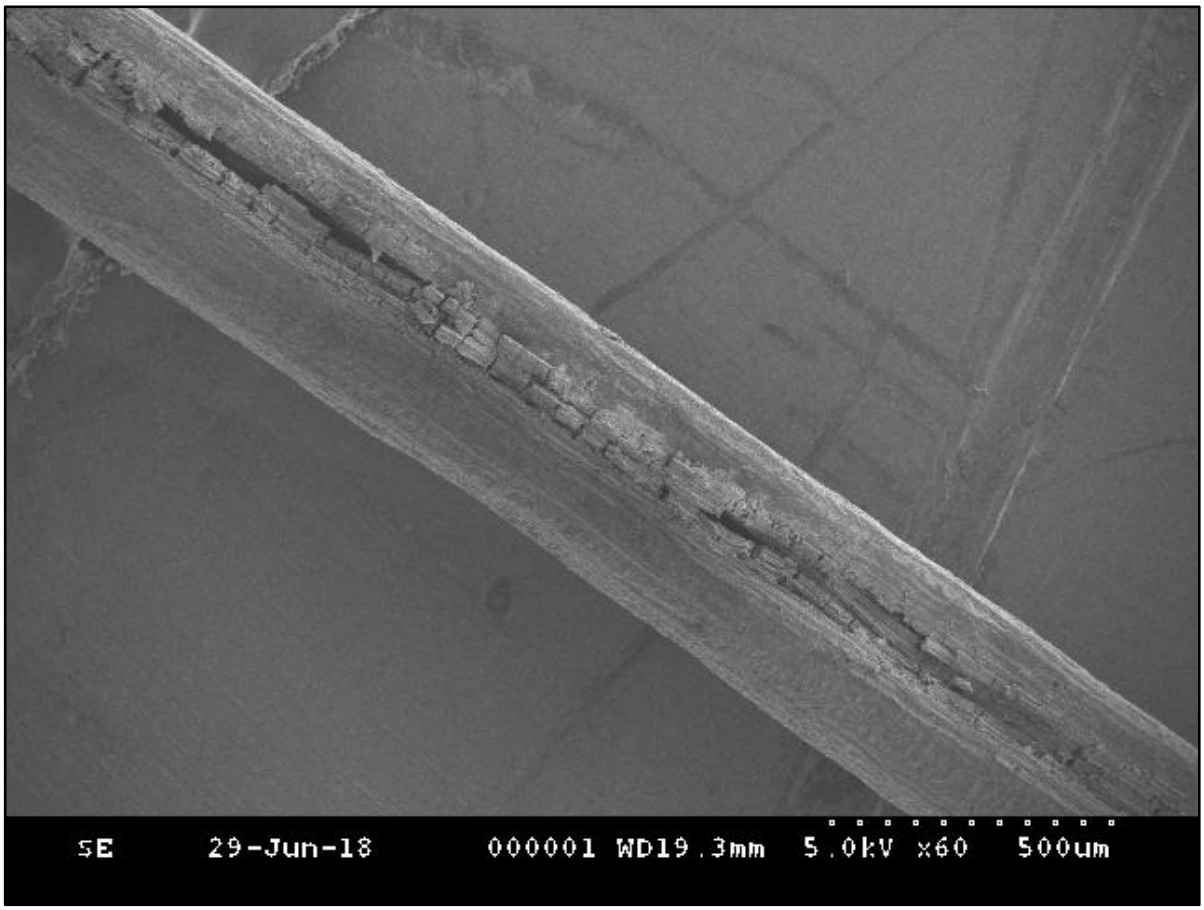


Figure 1 SEM micrograph of COIR fiber

C. Fabrication of coir fiber reinforced epoxy (CF-EP) composites

The CF-EP composites of different fiber lengths (10, 15, 20, and 25 mm) were fabricated using the hand-layup technique. Initially, the fibers were separated from pith, cleaned with distilled water, and dried in sunlight for 3 days to remove the entrapped water molecules prior to molding. The dried fibers were cut into a predefined length by using scissor and then thoroughly mixed with epoxy resin-hardener solution (2:1 by weight) to make the biocomposite sheets of dimensions 200 X 100 X 4 mm^3. The laminates were made by impregnating the randomly oriented chopped fibers with epoxy matrix. Table 2 records the designation and composition of the manufactured composites. The mold was made up of toughened glass of 6 mm thickness, covered with Teflon sheet and sprayed with a releasing agent to avoid the sticking of polymer resin. The compounded material

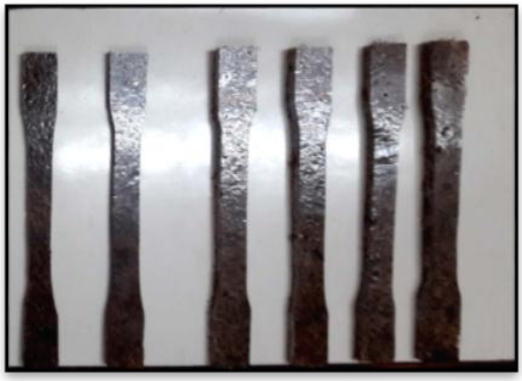
was poured into the mould [Fig. 2 (a)], closed the mould, and then pressed under 35 kg external loads for 24 hr curing at room temperature. The molded sheet, tensile specimens, and impact test specimens are shown in Figs. 2(b), 2(c), and 2(d) respectively.



(a)



(b)



(c)



(d)

Figure 2 (a) Glass mould with compounded materials (b) Coir fiber-Epoxy (CF-EP) biocomposite sheets (c) Tensile specimens (d) Impact test specimens

Table 2 Designation and composition of the coir fiber-epoxy (CF-EP) composites

Designation	Composition
EP	Epoxy (100 vol.%)
17(10)CF-EP	Epoxy (83 vol.%) + Coir fiber (17 vol.%) (fiber length of 10 mm)
17(15)CF-EP	Epoxy (83 vol.%) + Coir fiber (17 vol.%) (fiber length of 15 mm)
17(20)CF-EP	Epoxy (83 vol.%) + Coir fiber (17 vol.%) (fiber length of 20 mm)
17(25)CF-EP	Epoxy (83 vol.%) + Coir fiber (17 vol.%) (fiber length of 25 mm)
23(10)CF-EP	Epoxy (77 vol.%) + Coir fiber (23 vol.%) (fiber length of 10 mm)
23(15)CF-EP	Epoxy (77 vol.%) + Coir fiber (23 vol.%) (fiber length of 15 mm)
23(20)CF-EP	Epoxy (77 vol.%) + Coir fiber (23 vol.%) (fiber length of 20 mm)
23(25)CF-EP	Epoxy (77 vol.%) + Coir fiber (23 vol.%) (fiber length of 25 mm)
34(10)CF-EP	Epoxy (64 vol.%) + Coir fiber (34 vol.%) (fiber length of 10 mm)
34(15)CF-EP	Epoxy (64 vol.%) + Coir fiber (34 vol.%) (fiber length of 15 mm)
34(20)CF-EP	Epoxy (64 vol.%) + Coir fiber (34 vol.%) (fiber length of 20 mm)
34(25)CF-EP	Epoxy (64 vol.%) + Coir fiber (34 vol.%) (fiber length of 25 mm)
43(10)CF-EP	Epoxy (57 vol.%) + Coir fiber (43 vol.%) (fiber length of 10 mm)
43(15)CF-EP	Epoxy (57 vol.%) + Coir fiber (43 vol.%) (fiber length of 15 mm)
43(20)CF-EP	Epoxy (57 vol.%) + Coir fiber (43 vol.%) (fiber length of 20 mm)
43(25)CF-EP	Epoxy (57 vol.%) + Coir fiber (43 vol.%) (fiber length of 25 mm)

D. Mechanical Tests

The tensile and flexural (3 point bending) tests on CF-EP composites were carried out on Table Top Tinnus Olsen H50KS, Universal Testing Machine as per ASTM D 638 and 790 standard with a rate of 5 mm/min and 2 mm/min respectively. The impact strength of composite samples was determined by using an Izod impact tester as per ASTM D 256 standard. In all tests, the total four samples for each fiber length (10, 15, 20, and 25 mm) and volume content (17, 23, 34, and 43 vol.%) were tested and their average values were reported.

E. Water Absorption test

The water absorption test on epoxy matrix and the fabricated CF-EP composites was done in accordance with ASTM D 570 standard. Test was carried out in distilled water at room temperature, and the total absorbed water (%) for 24, 48, 72, 96, 120, 144, 168, 192, 216, 240, 264, 288, 312, 336, 360, 384, 408, 432, and 456 hrs was calculated by using the formula

$$\Delta \mathbf{w}(\mathbf{t})\% = \frac{w_t - w_o}{w_0} \times 100 \quad (1)$$

where $\Delta w(t)$ is total water absorption at time t, w_o and w_t are the initial dry weight and the wet specimen weight after time **t** respectively.

3. RESULTS AND DISCUSSION

All the fabricated biocomposites of varying fiber length and volume content were subjected to the mechanical and water absorption tests according to ASTM standard. The performance of composites during the tests is discussed below:

A. Tensile Properties

Figures 3-4 revealed the tensile properties of CF-EP and it was observed that the tensile strength was increased with the increase of fiber length and content up to 15 mm and 23 vol.% respectively. Further addition of fiber leads to the reduction in tensile strength and this diminution may be due to the agglomeration, poor dispersion, and insufficient wetting of the fibers with the polymeric resin. The maximum tensile strength of CF-EP composite is 19.23 MPa for 23% fiber volume content with 15 mm fiber length. Figure 3-4 revealed that the fiber length plays a significant role to affect the properties of composites. The tensile strength is less in case of long fibers (20-30 mm) and it was due to the folding, entanglement, and improper bonding of fibers with unfolded ones which results that they do not align properly in the longitudinal direction. Figure 4 depicts the effect of fiber length and content on Young's modulus of CF-EP composites and concluded on that the tensile modulus of CF-EP was increased with the increase of fiber volume fraction. The maximum tensile modulus (2.3 GPa) was obtained for 20 mm length with 34% fiber volume content. As the fiber content increases, gathering of the

fiber takes place which leads to the increase in stiffness of material.

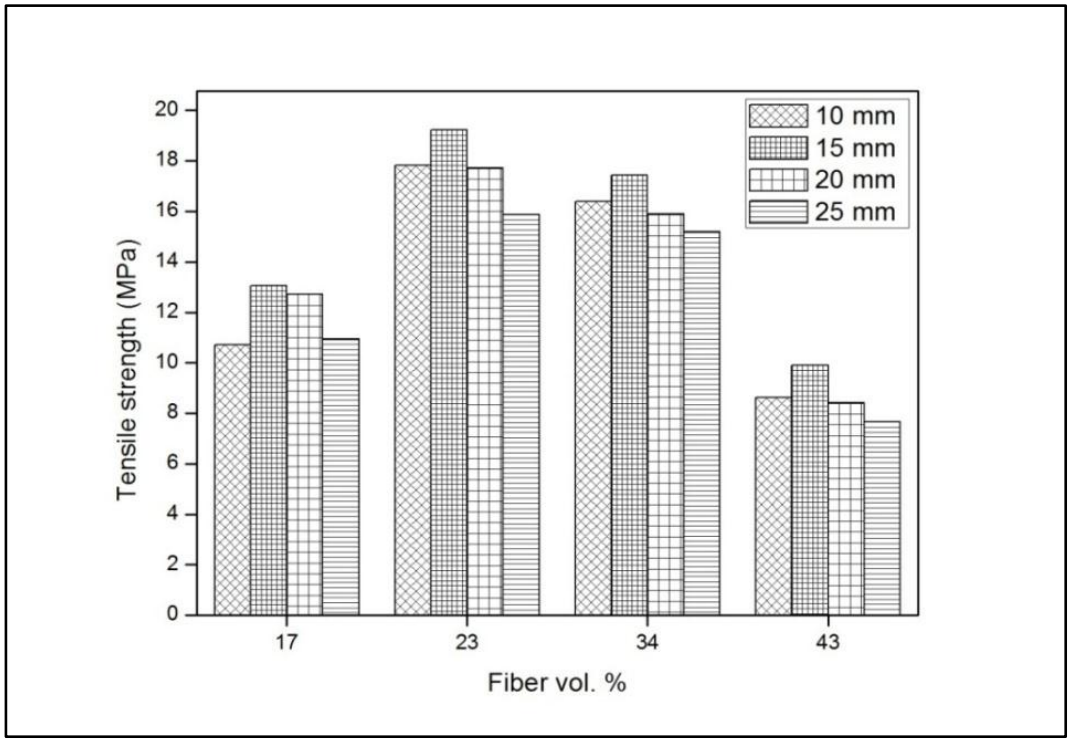


Figure 3 Effect of fiber length and volume content on tensile strength of CF-EP composites

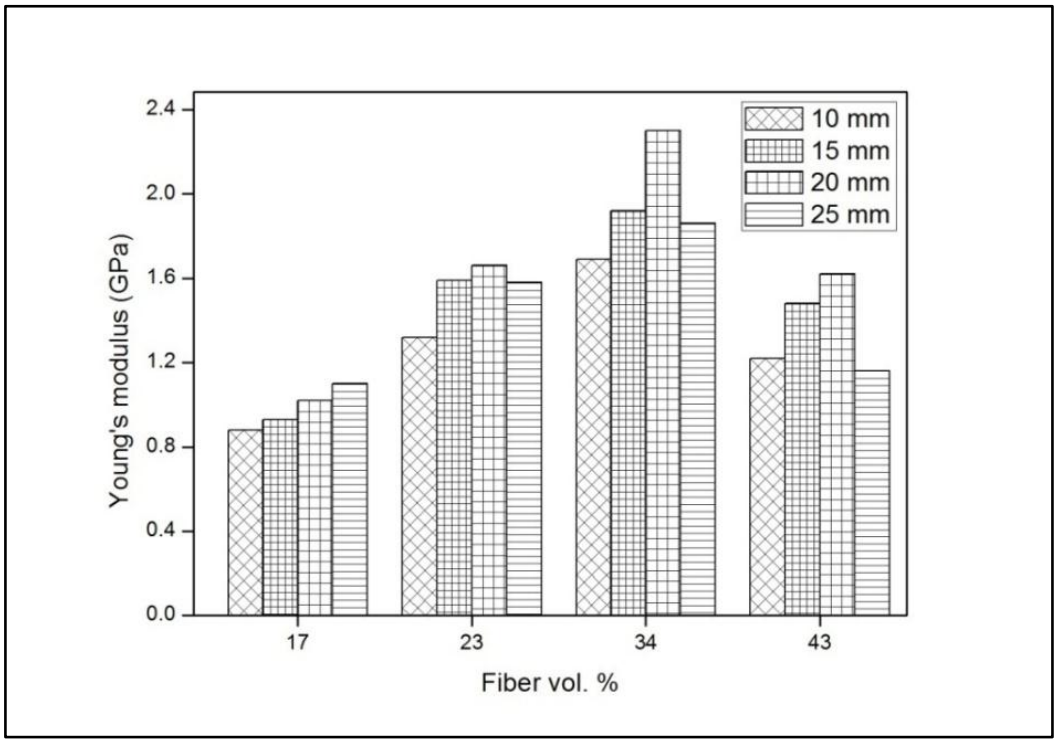


Figure 4 Effect of fiber length and volume content on Young's modulus of CF-EP composites

B. Flexural Properties

The significant effect of fiber length and content on flexural strength and modulus of CF-EP composites were recorded in Figures 5 and 6. It is worth to observe that the flexural strength of composites was increased with the increase of fiber length and volume content up to 20 mm and 23 vol.% respectively. Further increment in fiber length and content results the diminution in flexural strength and it may be attributed to the poor fiber-matrix interfacial bonding, curling of long fibers, increment in void density, and poor dispersion of fibers in the matrix resin. It was found that the 23 vol.% coir fiber epoxy composites exhibit higher flexural strength (i.e. 14.24%, 20.82%, 18.65%, and 20.95% for 10 mm, 15 mm, 20 mm, and 25 mm respectively) than the 17% CF-EP composites. The maximum flexural strength and modulus are 31.87 MPa and 2.41 GPa for 23(20)CF-EP and 34(20)CF-EP composites respectively.

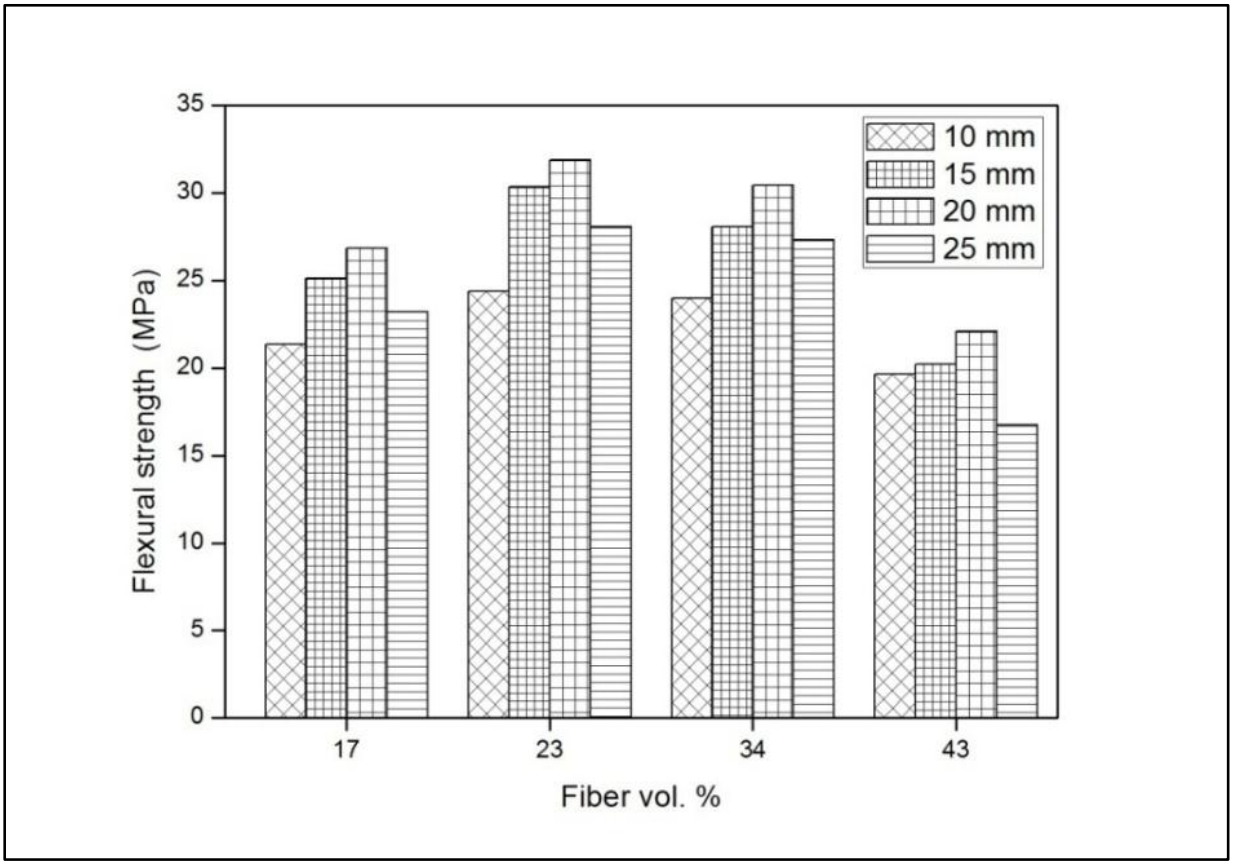


Figure 5 Effect of fiber length and volume content on flexural strength of CF-EP composites

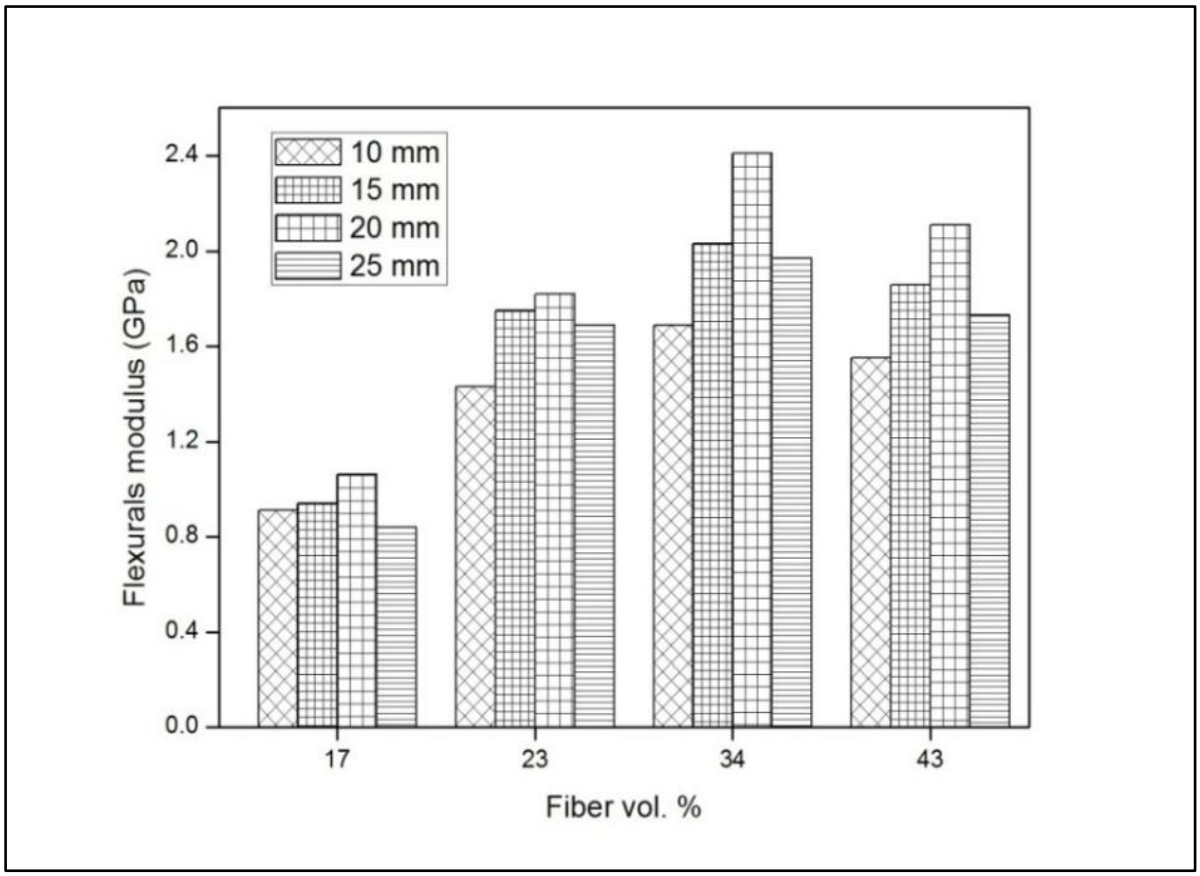


Figure 6 Effect of fiber length and volume content on flexural modulus of CF-EP composites

C. Impact strength

The effect of fiber length and content on impact strength of CF-EP composites are shown in Figure 7. From the obtained experimental data, it observed that the strength of CF-EP composites under impact loading (high strain rate) was increased with the increase of fiber volume fraction. It was due to the high energy dissipation in fibers pull-out process at higher loading of fiber. Generally, the impact strength of PMC's is highly dependent on the type of fiber, polymer, geometrical shape and size of fiber, and fiber-matrix interfacial bonding. Irrespective of fiber length, the maximum impact strength was reported with 43% fiber volume content. It is worth noting that the resistance to impact loading was increased with the increase of fiber length. Moreover, the composite consists of 20 mm long fibers with 43% fiber volume content possess maximum value of impact strength (11.33 kJ/m^2) than the other composites.

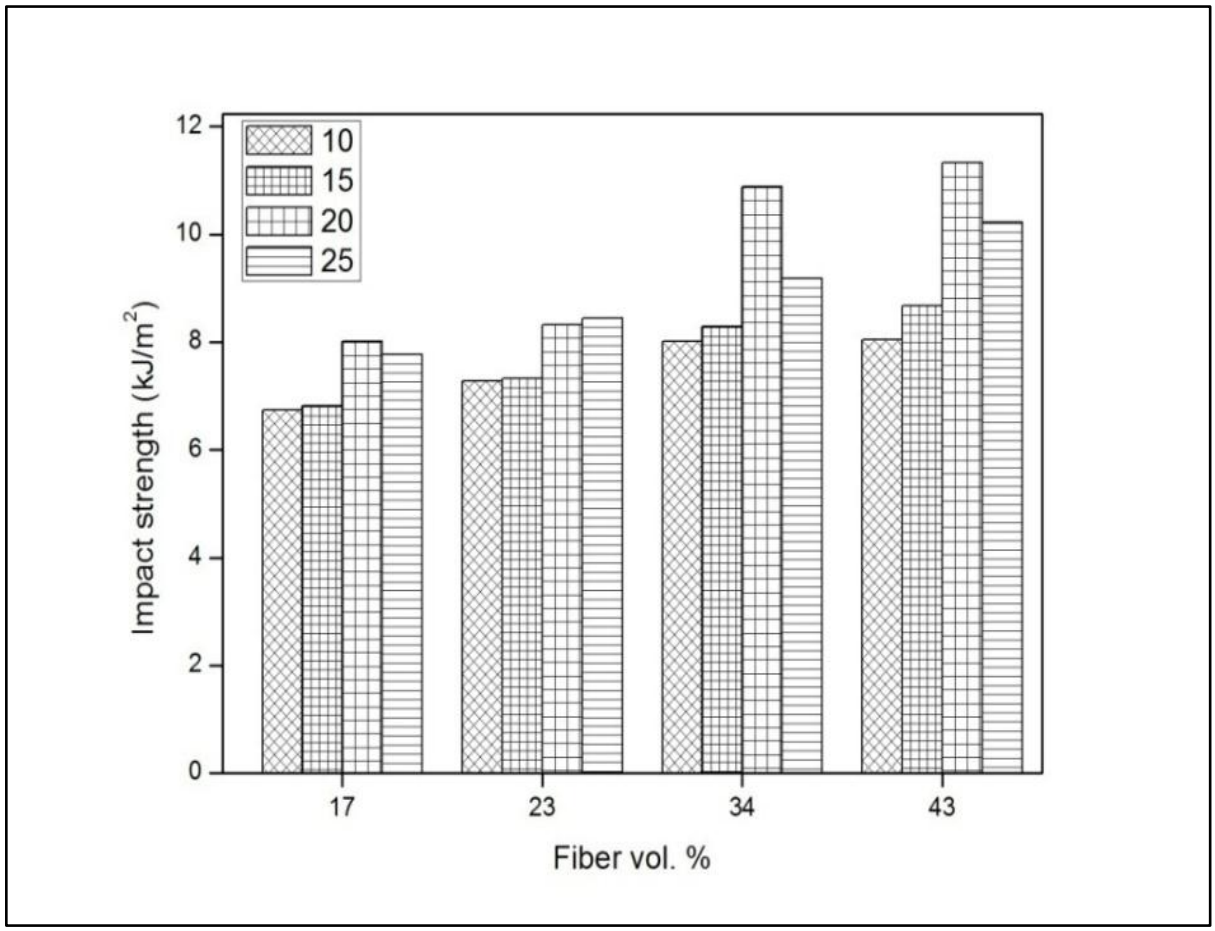


Figure 7 Effect of fiber length and volume content on impact strength of CF-EP composites

D. Water Absorption

The water uptake behavior of CF-EP composites with immersion time at different fiber length and content are presented in Figures 8-15. It observed that the water absorption was increased from 1.8% to 5.32% with the increase of fiber volume content from 17% to 43% respectively. This was due to the increase of hydroxyl (-OH) and other polar functional groups with the increase of fiber content. Previous works (G. Das *et al.* [8], G. Rajeshkumar *et al.* [9], V.K Bhagat *et al.* [12]) have also reported the similar results. The water absorption by all the samples increases monotonically with immersion time and becomes saturated after 360 hrs. The maximum amount of water absorption at the saturation level is tabulated in Table 3. From the experimental data, we can report that the variation in fiber length at constant value of fiber content not much influence the water absorption capacity of composites.

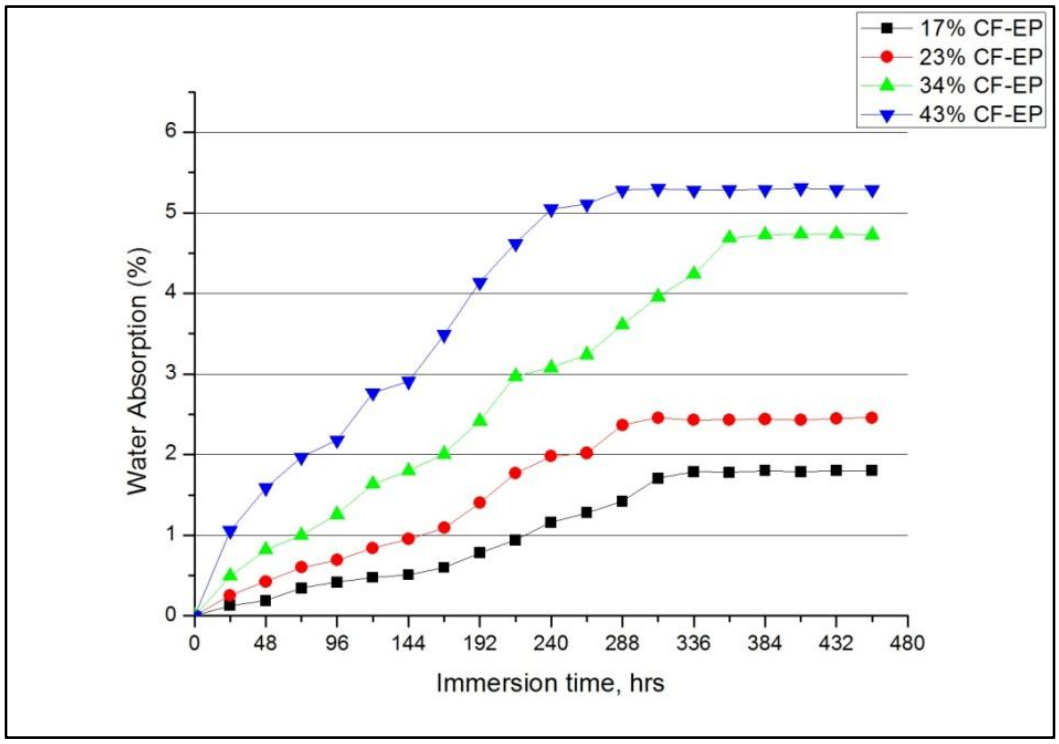


Figure 8 Effect of fiber content on water absorption of COIR/Epoxy composites; Fiber length – 10 mm

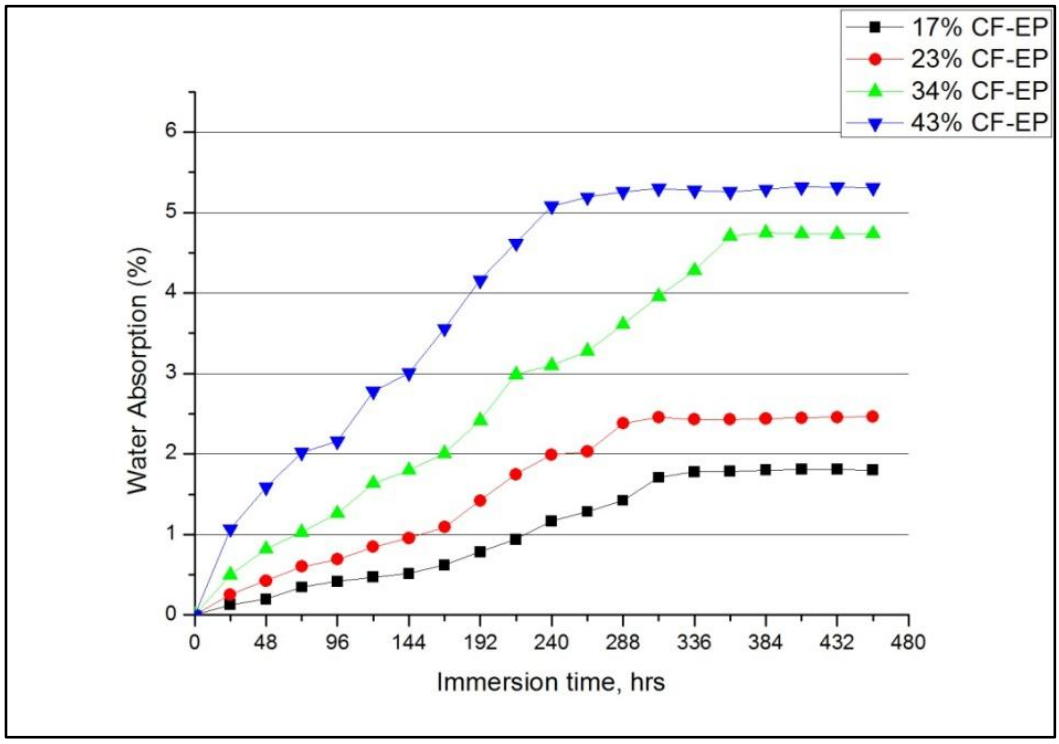


Figure 9 Effect of fiber content on water absorption of COIR/Epoxy composites; Fiber length – 15 mm

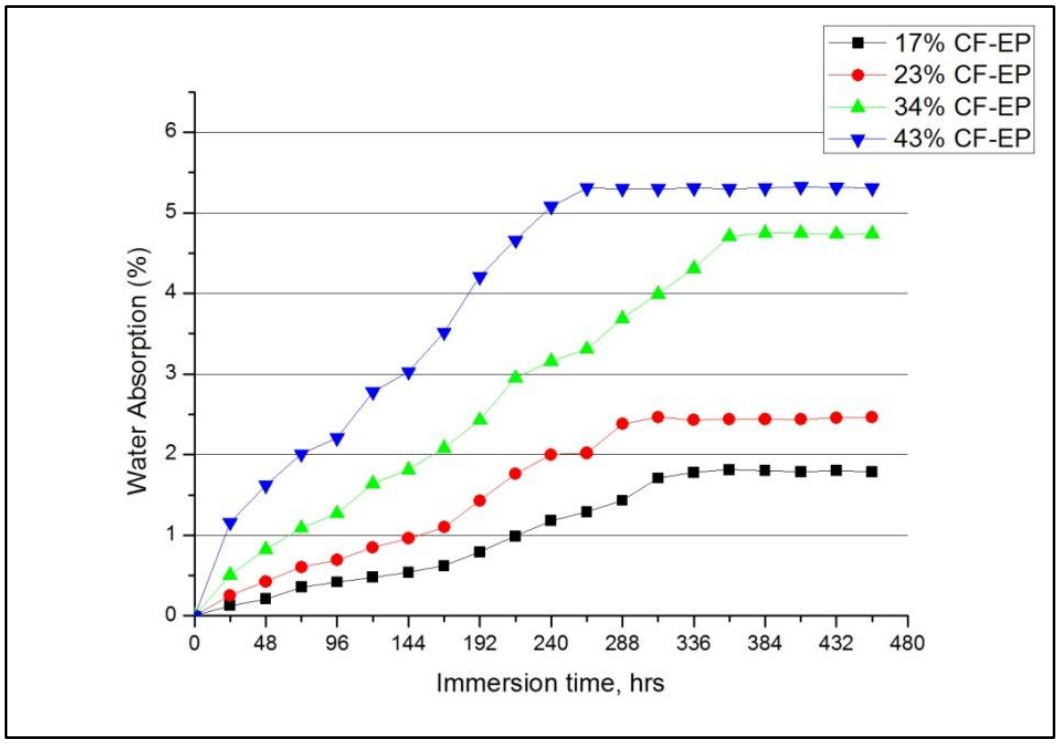


Figure 10 Effect of fiber content on water absorption of COIR/Epoxy composites; Fiber length – 20 mm

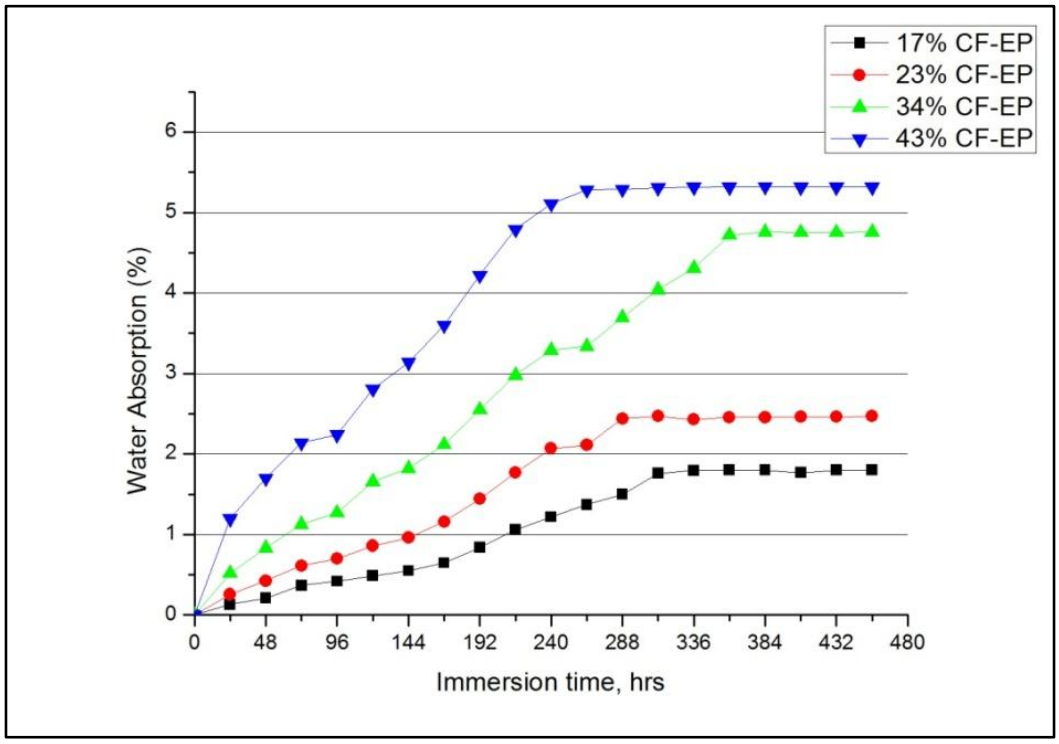


Figure 11 Effect of fiber content on water absorption of COIR/Epoxy composites; Fiber length – 25 mm

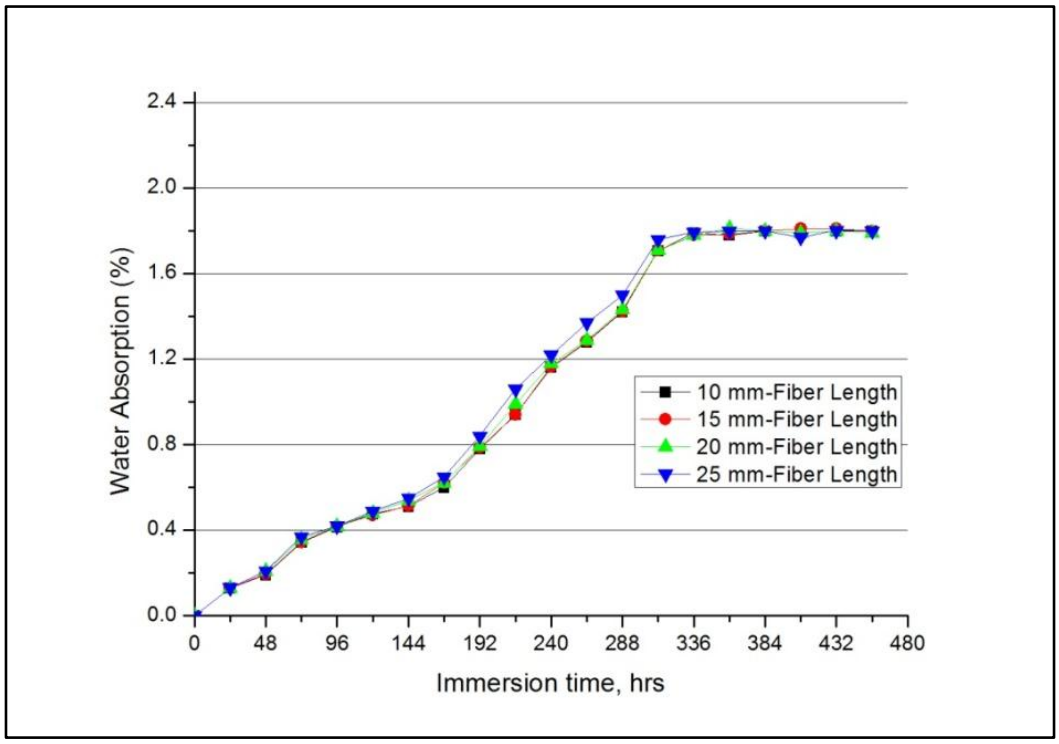


Figure 12 Effect of fiber length on water absorption of COIR/Epoxy composites; Fiber volume content – 17%

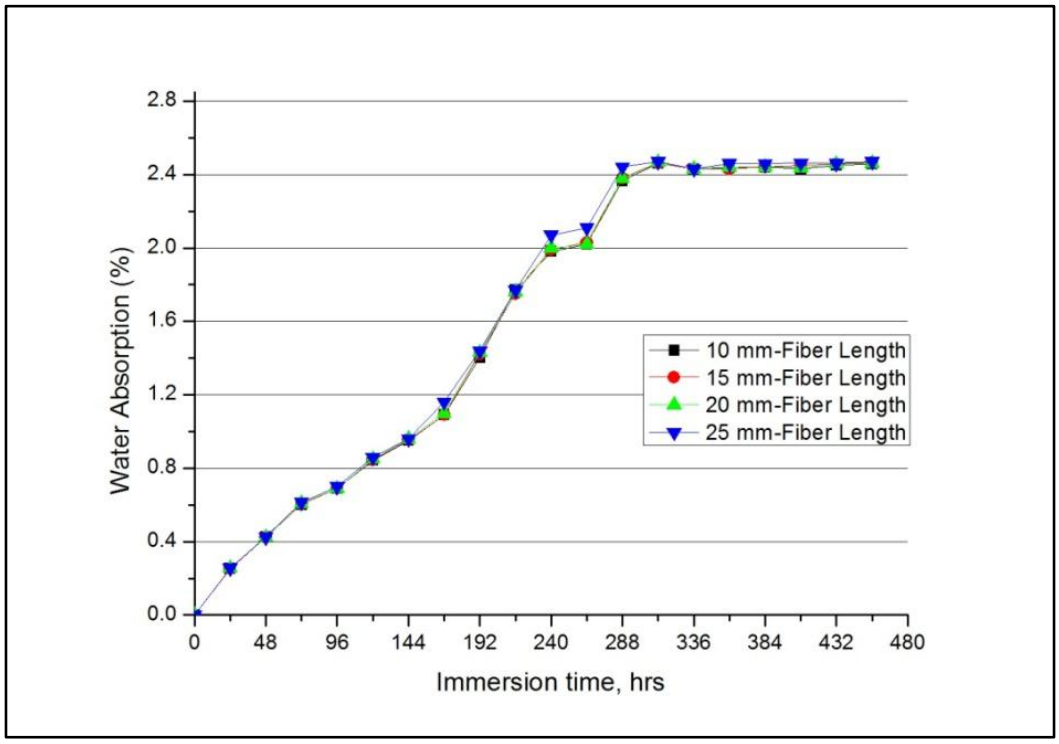


Figure 13 Effect of fiber length on water absorption of COIR/Epoxy composites; Fiber volume content – 23%

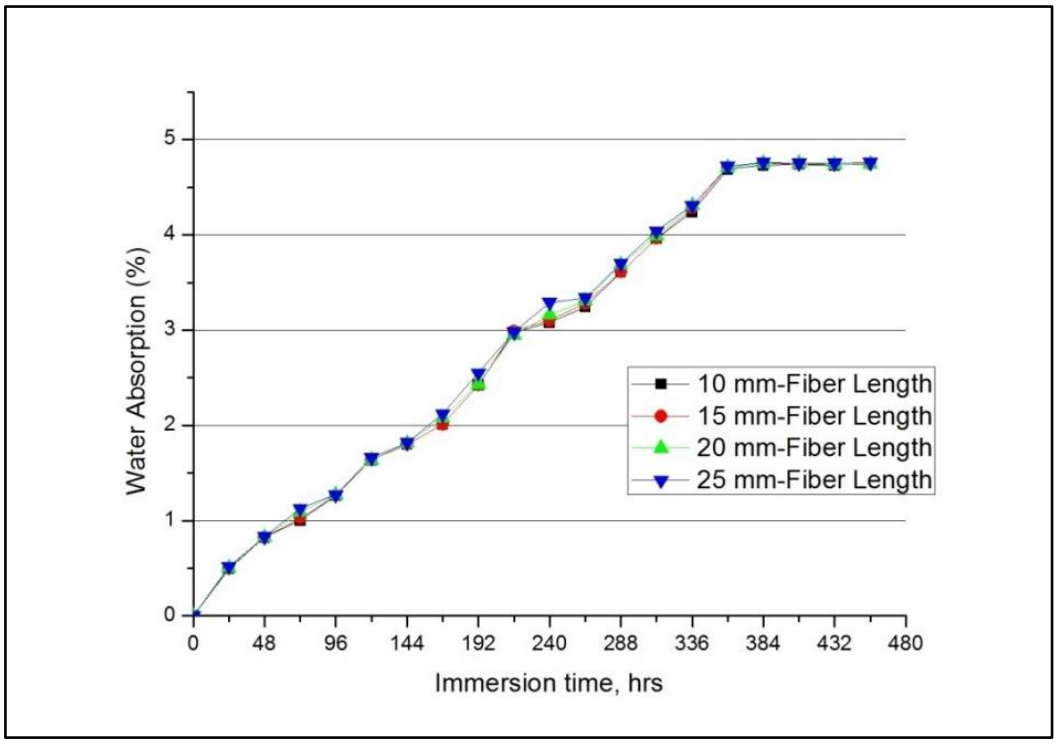


Figure 14 Effect of fiber length on water absorption of COIR/Epoxy composites; Fiber volume content – 34%

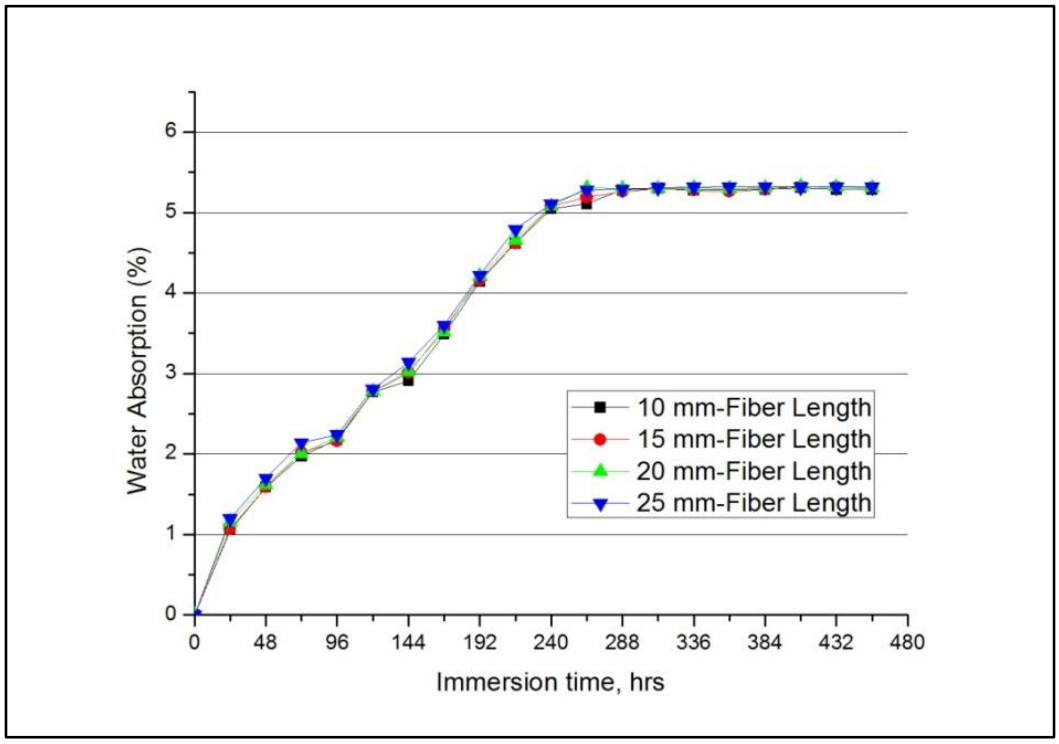


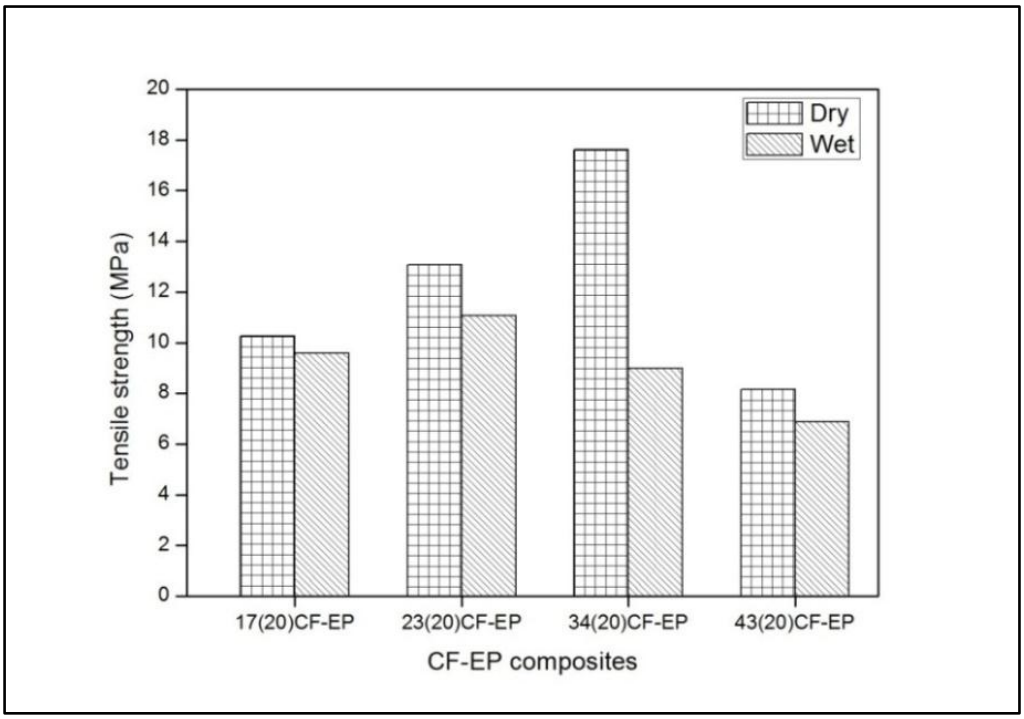
Figure 15 Effect of fiber length on water absorption of COIR/Epoxy composites; Fiber volume content – 43%

Table 3 Maximum water absorption for CF-EP composites

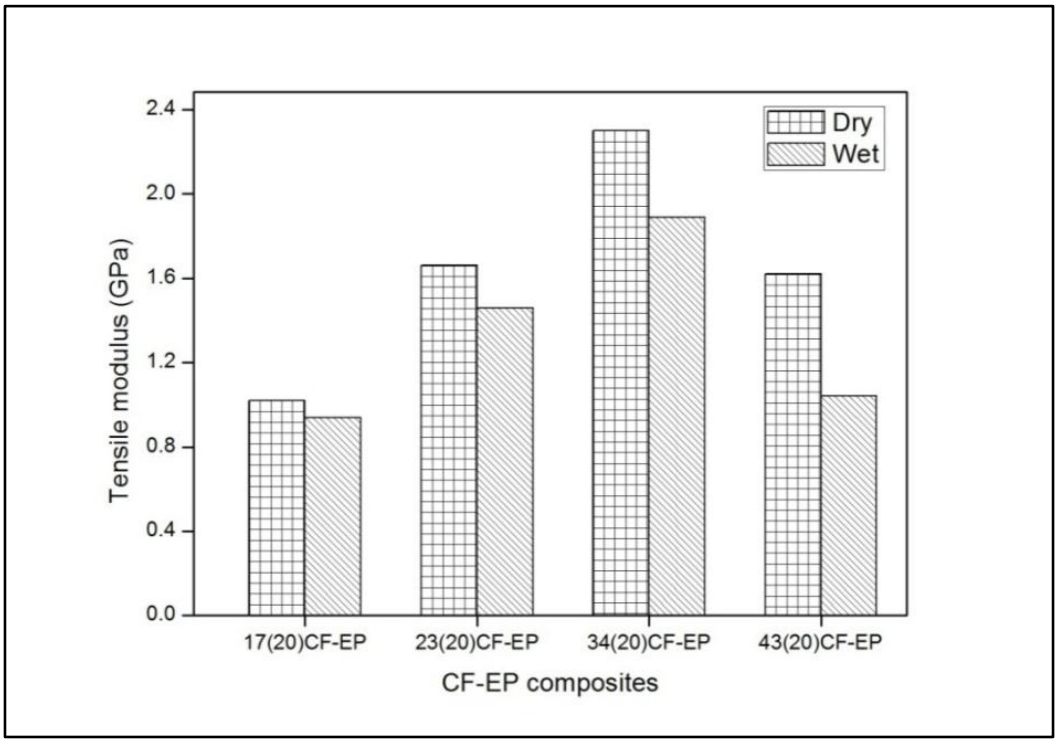
Fiber length	Coir fiber volume content (%)			
	17	**23**	**34**	**43**
10	1.80	2.45	4.74	5.31
15	1.81	2.46	4.75	5.32
20	1.81	2.46	4.75	5.32
25	1.80	2.47	4.76	5.32

E. Effect of water absorption on mechanical properties of CF-EP composites

The effect of water absorption on tensile, flexural, and impact properties of coir fiber reinforced epoxy composites was recorded in Figures 16, 17, and 18 respectively. It observed that the tensile and flexural properties (strength and modulus) of CF-EP composites were reduced after the immersion in water. The tensile strength of wet samples-17(20)CF-EP, 23(20)CF-EP, 34(20)CF-EP, and 43(20)CF-EP was found to be lower than the dry composite samples by 10.54%, 15.16%, 29.32%, and 37.02% respectively. This was attributed to the swelling of fiber, void formation, and poor stress transfer from matrix to the fibers. The flexural strength follows the same path as tensile strength and also showed the reduction in wet composites, 17(20)CF-EP, 23(20)CF-EP, 34(20)CF-EP, and 43(20)CF-EP by 8.45%, 8.75%, 10.31%, and 21.15% respectively than the dry composites of the same composition. Moreover, the tensile and flexural modulus of 43(20)CF-EP composite exhibits the maximum reduction, 35.8% and 24.64% respectively after immersion in water for 456 hrs. Figure 18 clearly revealed the improvement in impact strength of composites with the absorption of moisture. This may be due to the domination of fibers pull-out mechanism than the fiber fracture in wet composite samples which leads to the higher energy dissipation.

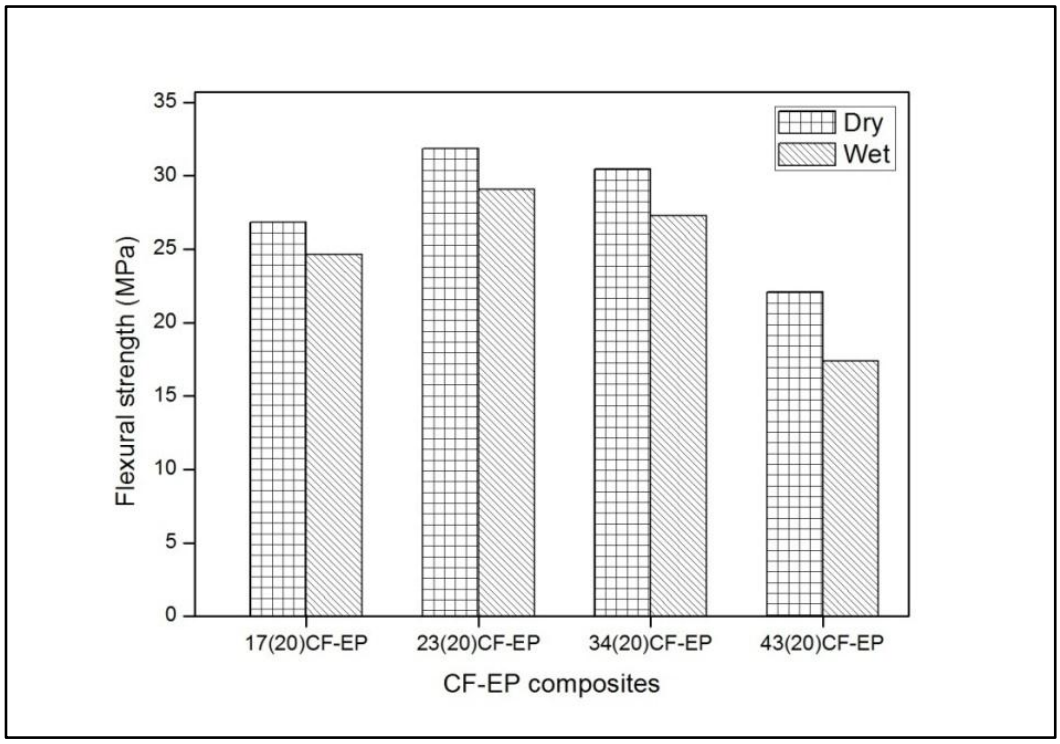


(a)

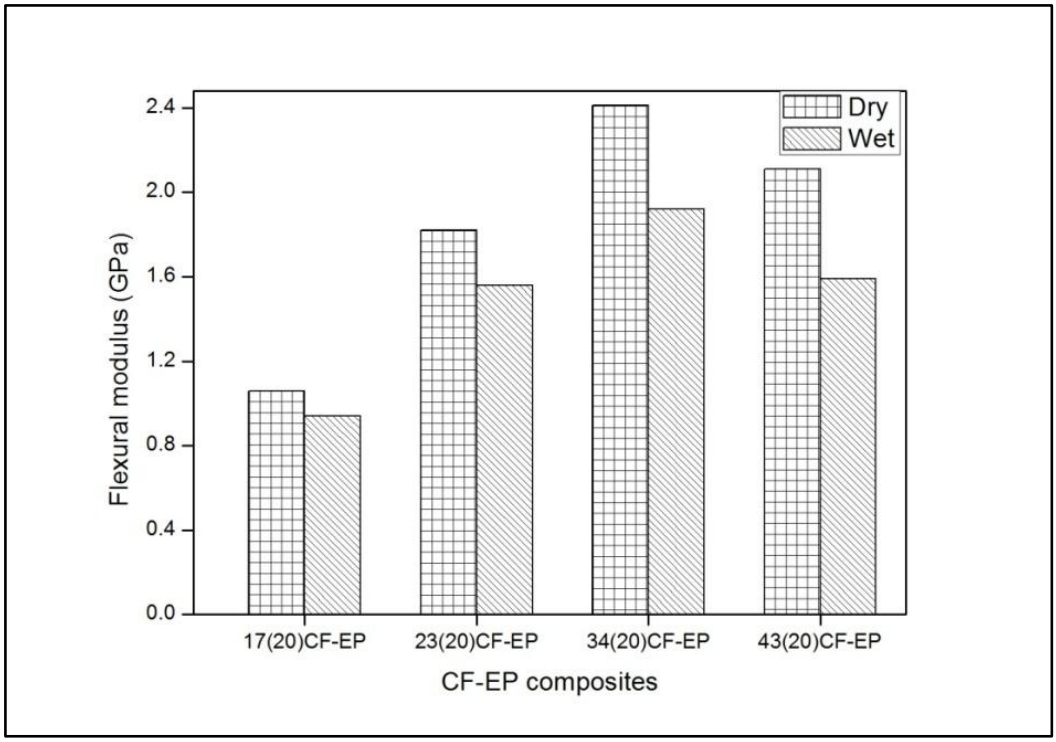


(b)

Figure 16 Effect of water absorption on (a) tensile strength (b) tensile modulus of CF-EP composites



(a)



(b)

Figure 17 Effect of water absorption on (a) flexural strength (b) flexural modulus of CF-EP composites

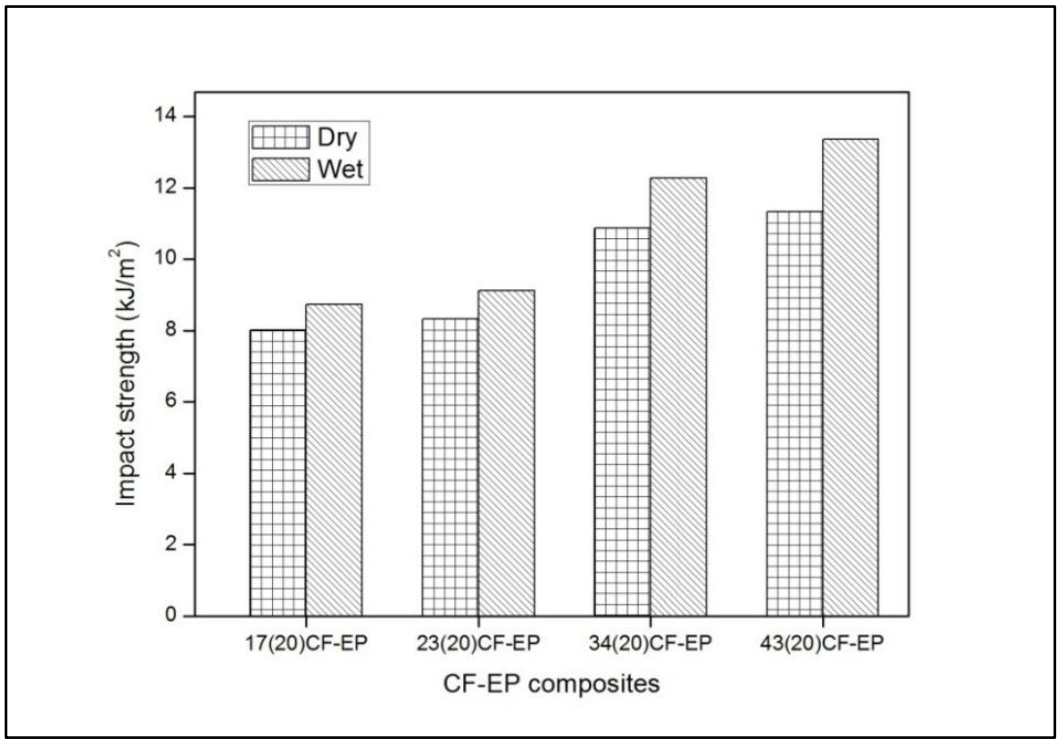


Figure 18 Effect of water absorption on impact strength of CF-EP composites

4. CONCLUSIONS

Based on the experimental results, following conclusions can be made:

- The composites reinforced with 23% coir fiber exhibit better tensile and flexural strength than the other composites. Beyond addition of fiber in a composite material leads to the improper wetting of fibers with resin.
- A coir fiber-epoxy composite loaded with 20 mm length of fiber possess maximum flexural strength, flexural modulus, tensile modulus, and impact strength. Further increase of fiber length leads to the increment in void density and poor dispersion of fibers in a matrix resin.
- The impact strength and moisture sorption affinity of composites increases with increase of fiber content. The maximum impact strength was obtained with 43% fiber volume content.
- The water uptake capacity of the composites is mainly affected by the fiber content than the fiber length.
- The tensile and flexural strength of wet composites are lower than the dry composites of the same composition.

REFERENCES

[1] N. Saba, M.T. Paridah, and M. Jawaid. "Mechanical properties of kenaf fiber reinforced polymer composite: A Review". Construction and Building Materials, Vol. 76, pp. 87-96, 2015.

[2] H.S. Kim, H.S. Yang, and H.J. Kim. "Biodegradability and mechanical properties of agro-flour-filled polybutylene succinate biocomposites". Journal of Applied Polymer Science, Vol. 97, pp. 1513-1521, 2005.

[3] S.H. Lee and S. Wang. "Biodegradable polymers/bamboo fiber biocomposite with bio-based coupling agent". Composites:Part A, Vol. 37, pp. 80-91, 2005.

[4] G. Romhany, J.K. Kocsis, and T. Czigany. "Tensile fracture and failure behaviour of thermoplastic starch with unidirectional and cross ply flax fiber reinforcements". Macromolecular Materials and Engineering, Vol. 288, pp. 699-707, 2003.

[5] T.T.L. Doan, S.L. Gao, and E. Mader. "Jute/Polypropylene composites I. Effect of matrix modification". Composite Science and Technology, Vol. 66, pp. 952-963, 2006.

[6] P.J. Franco and A. Gonzalez. "A study of the mechanical properties of short natural fiber reinforced composites". Composites: Part B, Vol. 36, pp. 597-608, 2005.

[7] D. Verma, P.C. Gope, A. Shandilya, A. Gupta, and M.K. Maheshwari. "Coir fiber reinforcement and appilications in polymer composites: A Review". Journal of Materials and Environment Science, Vol. 4(2), pp. 263-276, 2013.

[8] G. Das and S. Biswas. (2016). "Effect of fiber parameters on physical, mechanical, and water absorption behaviour of coir fiber-epoxy composites". Journal of Reinforced Plastics and Composites, Vol. 0(0), pp. 1-10, 2016.

[9] G. Rajeshkumar, V. Hariharan, TP. Satishkumar, V. Fiore, and T. Scalici. "Synergistic effect of fiber content and length on mechanical and water absorption behaviours of Phoenix sp. fiber-reinforced epoxy composites". Journal of Industrial Textiles, Vol. 0(00), pp. 1-22, 2016.

[10] R. Shrivastava, A. Telang, R. Rana, and R. Purohit. (2017). "Mechanical properties of Coir/G Glass fiber epoxy resin hybrid composite", Materials Today: Proceedings, 4: 3477-3483.

[11] S. Nallusamy, R. Rekha, and A. Karthikeyan. "Investigation on mechanical properties of coir fiber reinforced polymer resin composites saturated with different filling agents", IOP Conf. Series: Materials Science and Engineering, Vol. 225, pp. 1-9, 2017.

[12] V. Bhagat, S. Biswas, and J. Dehury. "Physical, mechanical, and water absorption behaviour of Coir/Glass fiber reinforced epoxy based hybrid composites", Polymer Composites, 2013 http://doi.org/10.1002/pc.22736.

[13] G. Goud and R.N. Rao. "Effect of fiber content and alkali treatment on

mechanical properties of *Roystonea regia*-reinforced epoxy partially biodegradable composites", Bulletin of Material Science, Vol. 34(7), pp. 1575-1581, 2011.

[14] M. Mittal and R. Chaudhary. (2018). "Effect of fiber content on thermal behavior and viscoelastic properties of PALF/Epoxy and COIR/Epoxy composites", Materials Research Express, 2018 http://doi.org/10.1088/2053-1591/aae274.

[15] U. Devi, SS Bhagawan, and S. Thomas. "Mechanical properties of pineapple leaf fiber reinforced polyester composites", Journal of Applied Polymer Science, Vol. 64, pp. 1739-1748, 1997.

[16] M.C.N. Yemele, A. Koubaa, A. Cloutier, P. Soulounganga, and M. Wolcott. "Effect of bark fiber content and size on the mechanical properties of Bark/HDPE composites", Composites: Part A, Vol. 41, pp. 131-137, 2010.

[17] N. Venkateshwaran, A. Elaya Perumal, and M.S. Jagatheeshwaran. "Effect of fiber length and fiber content on mechanical properties of banana fiber/Epoxy composite", Journal of Reinforced Plastics and Composites, Vol. 30(19), pp. 1621-1627, 2011.

[18] A. Sbiai, H. Kaddami, E. Fleury, A. Maazouz, F. Erchiqui, A. Koubaa, J. Soucy, and A. Dufresne. (2008). "Effect of the fiber size on the physicochemical and mechanical properties of composites of epoxy and date palm tree fibers", Macromolecular Materials and Engineering, Vol. 293, pp. 684-691, 2008.

[19] V.A. Alvarez, A.N. Fraga, and A. Vazquez. "Effects of the moisture and fiber content on the mechanical properties of biodegradable polymer-sisal fiber biocomposites", Journal of Applied Polymer Science, Vol. 91, pp. 4007-4016, 2003.

[20] Y.A. El-Shekeil, S.M. Sapuan, K. Abdan, and E.S. Zainudin. "Influence of fiber content on mechanical and thermal properties of kenaf fiber reinforced thermoplastic polyurethane composites", Materials and Design, Vol. 40, pp. 299-303, 2012.

[21] B. Bakri, S. Chandrabakty, and K. Putra. "Evaluation of mechanical properties of coir-angustifolia haw agave fiber reinforced hybrid epoxy composite", Jurnal Mekanikal, Vol. 8(1), pp. 679-685, 2017.

View publication stats

IOP Conference Series: Materials Science and Engineering

PAPER • **OPEN ACCESS**

EMISSION & PERFORMANCE ANALYSIS OF PRODUCER GAS AS IN IC ENGINE FUEL: A TECHNICAL REVIEW

To cite this article: Ankur Dixit *et al* 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **691** 012070

View the article online for updates and enhancements.

This content was downloaded from IP address 14.139.251.100 on 19/12/2019 at 04:49

EMISSION & PERFORMANCE ANALYSIS OF PRODUCER GAS AS IN IC ENGINE FUEL: A TECHNICAL REVIEW

Ankur Dixit[1], **Abhishek Srivastava**[2*], **Anmol**[2], **Akash Gupta**[2], **Nishant Garg**[2,] **Rajiv chaudhary**[3]

[1]Assistant Professor, Department of Mechanical Engineering, ABES Engineering College, Ghaziabad, India

[2]Student, Department of Mechanical Engineering, ABES Engineering College, Ghaziabad , India

[3]Professor, Department of Mechanical Engineering, Delhi Technological University, Delhi, India

*ankur.dixit@abes.ac.in

Abstract. For the sake of eco-friendly and large promise to be used in future producer gas plays an important role. It is prepared by gliding air and steam through the thick coal at different temperature range. A gasifier is used to gasify biomass such as rice husk, bagasse, wood chips. Due to inertness knocking tendency is higher in producer gas, use of producer gas tends to higher emission of CO, especially at lean condition. It results in rigorous environment damage. The use of producer as an alternative fuel is founded as the sustainable source of energy.

1. Introduction

As our country is the Biggest Producer of husk, straw etc. But still they are not utilized properly, so it is being used for making producer gas. In 1914 - 18 producer gas was introduced, but in world war second it was used widely in Europe, North America and Australia, producer gas was used in more than 1 million vehicles. Again the producer gas was seen in internal combustion engine in 1973 during the oil crisis. It is was not so much successful because during operation, the power should de-rated. The producer gas has low energy density as compared to other fuels like natural gas, diesel & gasoline (3,4). Producer gas is made by biogas's partial combustion. Such a process is known as Gasification.

Gasification is the process to convert the carbonaceous fuels into gaseous compound by providing less amount of oxygen which is necessary for the complete combustion of that fuel. By the conversion of fossil into CO, CO2, H. The output we get is Producer Gas. It is done at more than 700°C. Normally the temperature range is 1112°F-2732°F. It is used in dual mode engine in order to produced power and electricity. Gasification involves heating, pyrolysing partially oxidizing & reduction. The heating valve of output gas is found to be 950-1200KCa/m3. The calorific value of producer gas is very less.Composition of Producer Gas are:- CO (18-22%), H (9-13%), CH3(1-5%), HC(0.2-0.4%), CO2(9-12%), N(22-45%), H2O(4%).

There are several type of gasifier such as fluidized gasifier, fixed bed gasifier. There are three types, namely, up draft type, down draft type and cross draft type gasifier. Among the all, the down draft fixe bed gasifier is most preferable as there is less tar formation and need of cleaning is reduced.

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

2. Literature Review:

2.1. G.Sridhar [1]

Tested an engine at higher C.R. i.e. 17:1 for restrict the knock in the engine. Producer gas has calorific value 4.7-5.0 $MJNm^{-2}$. Variation in C.R. affects the engine's power output. Turbulent flame speed is the important factor affecting the rate of releasing the heat during process of combustion in the engine. The laminar speed of flame is decreased by 1/10 by increasing the initial pressure by 40 for the producer gas and air mixture. Growth in flame rate lead to increase in radicals released for C.R. 17:1, the output we get is 17.5 kW with overall efficiency 21%. It is compared with engine 21 kW with overall efficiency 31%. Overall efficiency means the shaft output per unit energy content. Efficiency and output is proportional to the compression ratio. At 18% efficiency, we get maximum output of 15.3 kW at C.R. 11.5:1.

2.2. Praveen Pandey [2]

Describes that producer gas made by rice husk replaced the diesel about 31% during operation in the dual fuel engine, the soot formation is much less when the proper supply of oxygen is given with that of producer gas. In gasifier we feed the waste products and after combustion of these product we get producer gas and then Scrubber, condenser, char and fabric filter are used to remove the Particulate impurities, water particles, moisture and fine dust respectively. After that the producer gas will supply to engine with diesel and then flue gas will test by Exhaust gas analyzer. On using diesel fuel and producer gas simultaneously then the opacity of exhaust gas increased and as the quantity of producer gas increases the opacity becomes maximum. If the less amount of nitrogen is present in biomass gasifier then the obtained producer gas which further mixed with diesel, the mixture has high calorific value. The parameters on which emission are measured such as HC, CO and smoke were larger in dual fuel type operation whereas in pure diesel these emissions are less.

2.3. Brenneisen et al [3]

Find 33.6% saving of fuel. On comparison of diesel fuel mode and dual fuel engine that is running at 1200 rpm it is founded that thermal efficiency increases from 31.92% to 34.85% and for 1400 rpm it is 31.03% to 35.60%. Temperature of gas at Exhaust is proportional to speed of the engine if the flow rate increases from 76 lpm to 125 lpm. The exhaust gas temperature also enhance from 17.99% to 46.03% with 1600 rpm. There is increment in smoke opacity from 4 to 6.07Km -1 with enhance the flow rate. The CO emission increased from 0.62% to 1.69% by volume at increasing flow rate from 76 lpm to 125 lpm due to sufficient oxygen it is founded that at 1600 rpm the CO level are minimum. Similarly HC emission also increases with increase in flow rate.

2.4. Jesper Ahrenfeldt et al. 2004 [4]

Describes the emission of CO is the degree of measurement of incomplete combustion of hydrocarbon and an indicator of PAH (poly nuclear aromatic hydrocarbon). This emission is due to less of oxidizer, pyrolysis of hydrocarbon results in PAH, which happens due to lack of oxygen. If temperature in boiler is high then PAH may decompose. Emission of CO from engine using producer gas is unlike the natural gas engine.CO emission are restricted by air-fuel ratio. It is characterized by the factor λ, which is the ratio of air-fuel (actual) and air-fuel ratio (stoichiometric) .For natural gas ($\lambda<1.2$), CO emission decreases due to partial oxidation of hydrocarbon if λ increases then CO emission also increases and for producer gas CO emission are higher essentially at $\lambda>1.5$.

Factors for PAH emissions are organic aerosol, higher hydrocarbon in fuel gas, engine lubricant when mixture is fuel rich that is $\lambda<1$. The hydrocarbons doesn't oxidize completely but get pyrolysis and results in formation of tar and PAH. PAH is also formed due to fuel trapped in cervices of the combustion chamber.

2.5. D.K. das et al. [5]
Describes that using gaseous fuel is founded as more advantageous because this given better efficiency, rate of release of heat is more, energy output can be controlled and managed. Producer gas needs 2.5-3 kg wood to make equal amount of energy as 1L of diesel. A downdraught type of gasifier was designed for 10.7 Nm^3/h with maximum load of 0.9 $N\text{-}m^3$ (cm^3-h). The diesel engine was analyzed on various load for 5 minutes each. Parameter were measured, the data is used to find thermal efficiency and specific diesel consumption in engine. The performance is evaluated with the help of above parameter. The thermal efficiency increased from 28% to 31% with increase in biomass moisture by 8% to 21%. The percentage of efficiency varies from material to material. The efficiency is much better with wood chips and corn cobs as compared to diesel mode and combustion is better with wood chips and corn cobs than pea stalks pigeon but energy content in pigeon pea stalk is high. There is 60%-64% less energy consumption for dual mode as compared to diesel mode.

2.6. Carlo Caliguiri and Massmiliano Renzi [6]
Analyzed the engine thermodynamically for sake of renewable sources, alternative methods were analyzed that results in introduction of producer in scenario. Although its quality is not satisfactory as compared to natural gas and gasoline still it is beneficial to environment as it is ecofriendly. SI engine needs modification to work with producer gas depends on ignition time, compression ratio of the engine and engine's rotational speed. Ignition time should be exceed by 30-40 degree with the help of low flame speed for CI engine de-rating is less than 30 percent for diesel engine by producer gas. So dual fuel engine comes in strategy. Dual fuel mode proves to be beneficial because of smooth operation with less noise. It is also economically fit and emission levels are very low. The analysis is done on thermodynamics' first law that is conservation of energy. More researches are done further on this topic.

2.7. G Sridhar et al. [7]
Worked on a open top reburn downdraft type of biomass gasifier which is beneficial for gasification purpose to produce less tar content gas, has a feasible calorific value [18,19]. The calorific value is about 5.2 ± 0.2 MJ/Nm^3 and there is a necessity of 1.2 to 1.4Kg of air which is needed for per Kg producer gas [17]. The analysis is done on dual fuel mode in CI engine. About 20-30% of power can be lossed when alone producer gas was used in the operation but the power can be compensated as toxic gases was generate in very less amount. NOx level is founded low in this as compared with diesel on the basis of lower peak flame temperature, but CO levels are more because of incomplete combustion. If we use producer gas only in SI engine the level of CO and NOx is founded less.

2.8. Vinay shrivastava et al. [8]
Describe that to design a downdraft type of gasifier was developed to running for 4-stroke, single cylinder. At different gas flow rates engine emission and performance characteristics were studies at various loads. In comparison the dual fuel type engine, brake thermal efficiency of diesel engine is more. Highest efficiency for diesel engine carryout was 21.5% whereas highest efficiency in dual fuel was 26%, 24.5%, & 25% for D+PG 4lpm, D+PG 6lpm & D+PG 9lpm respectively. At entire loading condition the specific energy consumption of diesel engine mode is constantly lower than that of dual fuel and the temperature of Exhaust Gas for diesel fuel mode is constantly less than dual fuel, this is because to provide extra energy to the engine . HC, CO, NOX exhaust are found that low compare to dual engine.

2.9. Felipe Centeno et al. [9]
Describe the performance on a model to check and developed the performance in a steady state of a downdraft type gasifier of biomass & spark ignition engine power system. In this paper they was used

fixed bed type downdraft gasifier & I.C engine of SI. The gasifier, divided into three parts for the calculation, namely pyrolysis, drying, oxidation and reduction sections. And further arithmetical model of gasifier consisting, three different sub-models, each processes describing the corresponding zone. And after further numerical results was obtained. Due to numerical results they get better overall efficiency in "biomass-to-energy" process of conversion.

2.10. S.Ramadhas, S.Jayaraj, C.Muraleedharan [10]

Describe the analysis of gasifier in this coir and wood chips used as the input. To study the work of gasifier and engine system, so engine is moving on various load and producer gas air flow ratios. At various loading condition engine with diesel are compared with that of dual fuel based engine on accounts of work and emission characteristics. After the performance we found that brake thermal efficiency of engine with diesel is more than dual engine. And in diesel mode the peak brake thermal efficiency is 25% was achieved. And in the operation in dual fuel mode, the maximal brake thermal efficiency are 21% at 70% of load with the use of wood chips and 19.9% at 70% load with coir-pith. The sp. energy consumption of diesel fuel mode is lower than dual mode at every loading conditions. In dual mode, there is decrease in the performance of engine with diesel or with rubber-seed oil acting as the pilot fuel. The amount of carbon monoxide emissions are higher with rubber seed oil-producer as compared to operation with diesel-producer gas below every loading conditions due to more consumption of fuel with less calorific value fuels. And it is also found that the amount of CO2 is higher in rubber oil-producer gas. At last, it is found that power generation cost is cheaper while using biomass as compared to the conventional power generation cost.

2.11. N.R. Banapurmath et al. [11]

Studied on honge oil and its methyl ester. Engine with dual fuel is used. To compare and evaluate the work of injection in CI engine which is connected with and without gas carburettor on dual fuel mode, to know the famous various other fuels like honge oil etc. Producer gas is the total alternative of fossil fuels. To find Honge oil methyl ester, so Tran's esterification is used to reduce the viscosity of Honge oil (Pongamia Pinnata Linn) which is to be selected according to the previous performance. We know the higher smoke emissions are produce by vegetable oils, so to improve their performance we adopted the dual fuel operation. Different values of brake thermal efficiency was obtained 23%, 22.25% and 24.25% with producer gas-Honge oil with methyl ester, producer gas with onl honge oil and producer gas-diesel, respectively. Emission such as NOx and smoke was reduce considerably in dual fuel operation. However, CO and HC emissions increased significantly.

2.12. Mohammad Asadullah [12]

Describe that one of the promising technologies is gasification is used to transform biomass into gaseous fuels energy generation. In this paper, they highlighted the biomass to electricity generation by removing the barrier in each and every steps. In this review, they discussed the effect of pre-treatment and parameters used in supply chain management, and refining and usability of gas for energy generation. According to previous paper, the improvement in method of gasification with effectively tar removing can efficiently elimination the consumption of biomass, and the biomass and logistics pre-treatment issues can be significantly reduced. With the help of small size (1-10MW) of the plant we can reduce the technical problems. Higher thermal efficiency is achieved due to combined agent of gasification, for instance, steam and air, was provided desirable gas composition for the engine.

2.13. P.Raman et al. [13]

Studied on to produce a clean gas to design a downdraft gasifier-dual system, used for electricity generation. With the the help of this system we improve a number of technical challenges. Indirect cooling of gas and dry gas purifying equipment is equipped with this system. Due to enhanced gasifier performance, the amount of tar present in the raw gas is not more than 100mgNm-3. Gasifier-dual

system produces the very lower tar content, are 67mgNm-3 against 711mgNm-3 of the reactor. So, low dust content gas is produce with improved system, are 53mgNm-3 against 1360mgNm-3 at the end of the reactor. Level of tar content is 35mgNm-3 at the end of the gas purifying train. Gas calorific value is 5.3MJNm-3. Fuel wood/kg a producer gas of specific gasification rate is 2.8Nm-3. The normal value of fuel consumption is 1.5-1.6 Kg kWh-1 against the 1.1 kg fuel wood/kWh. Efficiency of an electric transformation of the biomass system is noticed to be 21%.

3. Performance parameter:

3.1. Brake thermal efficiency:

The energy in brake power per unit input energy is known as brake thermal efficiency.

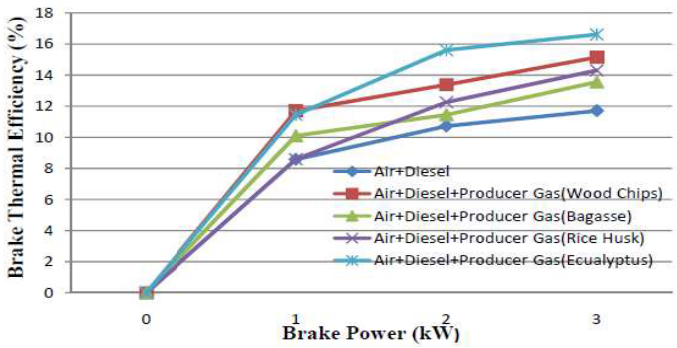


Fig.1 Brake thermal efficiency (%) vs. Brake Power (kW) [20]

Brake thermal efficiency get increased by 4% by the use of producer gas. The above figures indicates the variability of brake thermal efficiency along with the brake power for different types of fuels. The curve of rice husk is varying same as the air + diesel curve up to 1kW brake power bur after that there is increase in brake thermal efficiency of the fuel with rice husk. Similarly, fuel with eucalyptus have same efficiency as wood chip's fuel but the efficiency of eucalyptus based fuel is increased after 1kW of brake power. In this way the above curves shows the variability of efficiency at different value of brake power.

3.2. Exhaust gas temperature

It is the indication of how energy is converted in work.it is founded higher in leaner mixture. If the exhaust gas temperature is higher than NOx level also increases.

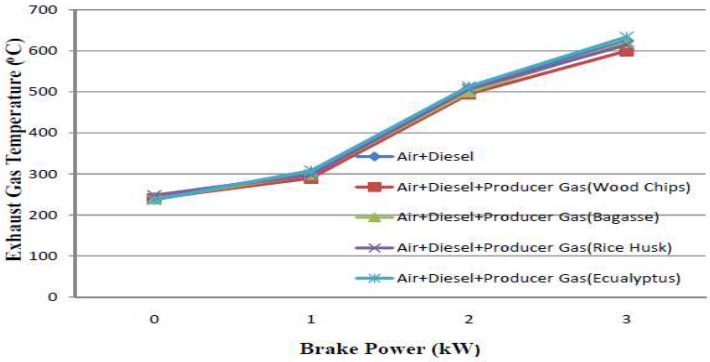


Fig.2 Exhaust Gas Temperature (°C) vs. Brake Power (kW) [20]

The above indicates the variability of exhaust gas temperature along with the brake power on various types of mixtures of fuels.

3.3. Brake specific energy consumption

It is termed as the fuel consumed/brake power.

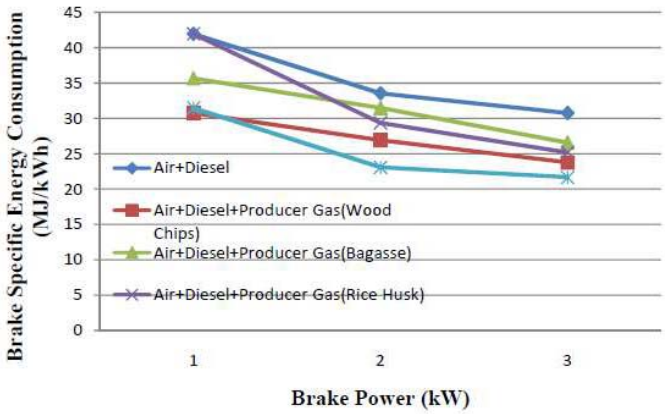


Fig.3 Brake Specific Energy Consumption (MJ/kWh) vs. Brake Power (kW) [20]

In the above fig. it is shown that how the Brake Specific Energy Consumption changes at different value of brake power for different types of fuels such as Air+Diesel, Air+Diesel+producer gas(wood chips), Air+Diesel+producer gas(Bagasse) & Air+Diesel+producer gas (Rice husk). In which rice husk based shows better results than other fuel.

4. Conclusions

From the above paper various things are concluded that for CR17:1 overall efficiency for 17.5kW is 21% and for 21kW this is found to be 31%. With 1200rpm, diesel mode and dual fuel mode engine gives 3% increment in thermal efficiency and for 1400rpm, it is increased by 4%. If we increased flow rate then emission of CO and HC is also increased but by providing sufficient oxygen, CO emission can be reduced. Producer gas needs 2-3kg to give equal amount of energy as 1L diesel. Flow rate also enhance the saving of fuel.

5. References

[1] G. Sridhar, P.J. Paul, H.S. Mukunda. Biomass dervied producer gas as a reciprocating engine fuel- an experimental analysis, 2001.

[2] Praveen Pandey. Energy and Emission analysis of dual fuel CI Engine Using Diesel and Biomass Derived Producer Gas, March 2010.

[3] S.Dasappa & H.V. Sridhar. Performance of a diesel engine in a dual fuel mode producer gas for electricity power generation. 20 Sep 2011.

[4] Jesper Ahrenfeldt, Torben Kvist Jensen, Ulrik Henriksen , Jesper Schramm and Benny Gobel, CO Emission from Gas Engine Operating on Biomass Producer Gas, 2004.

[5] D. K. Das, S. P. Dash, M. K. Ghosal, Performance evaluation of a diesel engine by using producer gas from some under-utilized biomass on dual-fuel mode of diesel cum producer gas, 2012.

[6] Carlo Caligiuri, Massimiliano Renzi, Combustion modelling of a dual fuel diesel – producer gas Compression ignition engine, August 2017.

[7] G Sridhar, S Dasappa, H V Sridhar, P J Paul, N K S Rajan, Gaseous Emissions Using Producer Gas as Fuel in Reciprocating Engines, April 2005

[8] Vinay shrivastava ―design and development of downdraft gasifier for operating ci engine on dual fuel mode‖ department of mechanical engineering National institute of technology Rourkela may 2012 1-78

[9] Felipe Centeno, Khamid Mahkamov, Electo E. Silva Lora, Rubenildo V. Andrade ―Theoretical and experimental investigations of a downdraft biomass gasifier - spark ignition engine power system‖ Renewable Energy 37 (2012) 97-108

[10] A.S. Ramadha , S. Jayaraj, C. Muraleedharan ―Power generation using coir-pith and wood derived Producer gas in diesel engines‖ Fuel Processing Technology 87 (2006) 849–853

[11] N.R. Banapurmath, P.G. Tewari ―Comparative performance studies of a 4-stroke CI engine operated on dual fuel mode with producer gas and Honge oil and its methyl ester (HOME) with and without carburetor‖ Renewable Energy 34 (2009) 1009–1015

[12] Mohammad Asadullah ―Barriers of commercial power regeneration using biomass gasification gas‖ Renewable and Sustainable Energy Reviews 29(2014)201-215

[13] P. Raman, N.K. Ram, Ruchi Gupta ―A dual fired downdraft gasifier system to produce cleaner gas for power generation‖ Energy 54 (2013) 302 to 314

[14] Rohan R. Patel, DR. Rajesh C. Iyer, Hiren Mahida & Nirmalkumar, A Review paper on Theoretical & Experimental Investigation of a Dedicated Producer Gas Engine Technology, 2015.

[15] Mahida Hirenkumar Ranjitsinh, Rajesh C. Iyer, A Techniacl Review of Producer Gas as an Alternative Fuel for Internal combustion Engine, 1 January 2015.

[16] Heywood John B., (1989), Internal Combustion Engine Fundamentals, International Edition, McGraw-Hill, New York, pp. 572-589.

[17] Martin J. and Wauters P. Performance of Charcoal Gas Internal Combustion Engines, Proceedings of International Conference- New Energy Conversion Technologies and Their Commercialization, Vol 2, PP 1415-1424.

[18] Ramachandran A. (1993) Performance studies on a wood gas run IC engine, Procedings of Fourth National Meet on Biomass Gasification and Combustion, Mysore, India Vol 4, PP 213-218.

[19] Francisco V. Tinaut, Andres Melgar, Alfonso Horrilo, Ana Diez de la Rosa (2006): Methods of predicting the performance of an internal combustion engine fuelled by producer gas and other low heating value gas, Fuel Processing Technology 87, PP 135-142.

[20] J. P. Yadav, Performance Analysis of Producer Gas Based Diesel Engine, Volume 2, Issue 1, and ISSN: 2277 – 5668

AIP Conference Proceedings

Entropy as a measure of implied volatility in options market

Cite as: AIP Conference Proceedings **2183**, 110005 (2019); https://doi.org/10.1063/1.5136222
Published Online: 06 December 2019

H. C. Taneja, Luckshay Batra, and Pulkit Gaur

View Online Export Citation

Lock-in Amplifiers up to 600 MHz
Zurich Instruments
Watch



© 2019 Author(s).

Entropy as a measure of implied volatility in Options Market

H.C. Taneja[a)], Luckshay Batra[b)] and Pulkit Gaur[c)]

Department of Applied Mathematics, Delhi Technological University, Delhi-110042, India

[a)]hctaneja@dce.ac.in
[b)]luckshay25@gmail.com
[c)]pulkitgaur_2k14@dtu.ac.in

Abstract. This paper's main contribution is to liken various techniques, one based on the standard deviation, a statistic and one based on the investor's expectation on the future movements of the underlying asset price (implied) and the last, centred on the concept of Shannon entropy. The empirical analysis is carried out so as to find some relationship between the three different approaches.
Keywords: Shannon entropy, stock market volatility, implied volatility, financial markets.
PACS: 89.65.Gh, 89.70.Cf

INTRODUCTION

There has been a growing debate on the stock market volatility over the past few years [1]. Shiller [2] asserted that the measured volatility of the stock market was erratic mostly with the predictions of models based on current values. Schwert [3] brought a new insight into this, who asked "Why does stock market volatility change over time?".

We use entropy to measure the uncertainty in the Indian financial markets from 2000-2017 since the standard deviation can discern only linear relationships. It is also worth investigating deeper into the working of these entropic measures in the situations of financial turmoil. The paper is organised as follows. In Section 2, the important aspects of entropy and volatility are presented. In Section 3, the empirical study is presented and the paper ends with a conclusion in Section 4.

Theoretical Framework: Volatility and Entropy

In this section, first we explain the concept of volatility [4] and then we take a closer look at information theoretic measure i.e., entropy.

Volatility: *Historical Volatility* is the annualized volatility σ, the standard deviation of the annual logarithmic returns of an instrument, defined as $\sigma_T = s\sqrt{\Delta T}$ where T stands for trading days per year (usually 252), s is the standard deviations of stock price returns over time. Then, σ gives us an estimate of the historical volatility of the stock. The stock price S_i is usually observed daily over a period of time to calculate the historical volatility of a stock price, and and then we let $u_i = ln\left(\frac{S_i}{S_{i-1}}\right)$ for $i = 0, 1, 2, ..., n$ and the standard deviation s of $u_i's$ is given as $s = \sqrt{\frac{1}{n-1}\sum(u_i - \bar{u})^2}$ where $\bar{u}$ is the mean of u_i. The volatility is given by $\sigma\sqrt{T}$, and since s is an estimate of σ we can write volatility as $\hat{\sigma} = s\sqrt{T}$.

On the other hand, *Implied volatilities* are the volatilities resulting from the market observed option prices. It is used to monitor the opinion of the market or a trader about the volatility of the security. Just like market as a whole, implied volatility is subject to variability. The implied volatility is the value of σ when substituted into the Black-Scholes-Merton formula [5] for the European call option price: $C = S_0N(d_1) - Ke^{-rT}N(d_2)$ where, N(.) is a cumulative distribution function for a standardized normal random variable, i.e., $0 < N(x) < 1$, and r and σ are kept

Third International Conference of Mathematical Sciences (ICMS 2019)
AIP Conf. Proc. 2183, 110005-1–110005-4; https://doi.org/10.1063/1.5136222
Published by AIP Publishing. 978-0-7354-1930-8/$30.00

as constant, given by $N(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp^{-\frac{1}{2}y^2}\, dy$ where, $d_1 = \frac{\ln(\frac{S_0}{K}) + (r + \frac{1}{2}\sigma^2)T}{\sigma\sqrt{T}}$ and $d_2 = \frac{\ln(\frac{S_0}{K}) + (r - \frac{1}{2}\sigma^2)T}{\sigma\sqrt{T}}$ gives the value of C. But its not possible to invert the above equation and express σ as an explicit function of the other parameters. Hence, we have to use some numerical methods like Bisection Method or Newton Raphson Method to find this volatility.

Entropy: The concept of entropy in the communication system given by Shannon [6] is an alternative way of studying volatility in the stock market. This can be used to describe the volatility's nonlinear dynamics. For the given probability distribution of a random variable X, $p_i = p(x_i = i)$, $i = 1, 2..., n$, the Shannon entropy S is

$$S(X) = -\sum_{i=1}^{n} p_i ln p_i\ ,\ \ 0 \le p_i \le 1\ and\ \sum_{i=1}^{n} p_i = 1 \tag{1}$$

The units of entropy is bits (when base is 2) and nats (when base is e). The properties of entropy are well established as a measure of uncertainty.

Empirical Analysis

This section examines the objective of the study to test whether entropy is a good measure for volatility in stock returns and also to compare historical volatility with implied volatility.

In our empirical study[1], we have used the closing price data[2] from Jan 2000 to Dec 2017 of top 10 Indian companies: Tata Consultancy Services (TCS), Hindustan Unilever Limited (HUL), India Tobacco Company (ITC), Oil and Natural Gas Corporation Limited (ONGC), Infosys, Reliance, Maruti Suzuki, ICICI Bank, HDFC Bank, State Bank of India (SBI); top four public sector banks: Bank of Baroda, Punjab National Bank (PNB), IDBI Bank, State Bank of India (SBI); and top four private sector banks: Yes Bank, Kotak Bank, HDFC Bank, Axis Bank in terms of their market cap in the year 2018 which has 4478 observations to make our analysis meaningful. The reason of choosing this period was that the period has several scenarios of markets in it, the crash due to the economic crisis as well as the sharp rise due to stimulus package given by the Indian government for the revival.
The data has daily closing prices of the index which was used to calculate the daily returns. With the help of these daily returns, we calculate the annualized historical volatility. We compute a time series of historical volatility by taking a calculation window of 1 year and the time step is 1 day. As we constructed a time series for historical volatility, we did the same for entropic measure i.e., Shannon entropy so as to compare them and visualize them correctly. The entropy was calculated using the natural logarithm and the unit for entropy here is nats.

After analyzing the time series, we find that entropy moves in a similar way as of volatility, see Figure 1 (for Indian companies), Figure 2 (for Indian public sector banks) and Figure 3 (for Indian private sector banks), and it has been seen that in periods of crash or jumps (highly volatile), volatility and entropy show similar signals but in period of stability, entropy doesn't give much information about the uncertainty. We have built correlation matrices for sector-wise analysis to check the relation in between entropy, historical volatility and implied volatility, see Table 2, Table 4 and Table **??**. We can also say that the entropy is mean reverting similar to volatility and hence it can be used in mean reversion trading strategies.

Next, our concern is to model Implied Volatility (IV) as a function of Historical Volatility (HV) and Entropy (S). As per Gulko [7], the entropy based IV subsumes all the information about historical volatility and entropy. So, we try to make a model as:

$$IV = aHV + bS + c \tag{2}$$

As per our above model, using the data from Table 1, Table 3, and Table **??**, we calculated $a = -0.05$, $b = 41.46$, $c = -147.93$ by using multiple regression and from here we observe that variation in S effects IV much more as compared to the variation in HV. So entropy can be seen as a suitable alternate to implied volatility in financial markets.

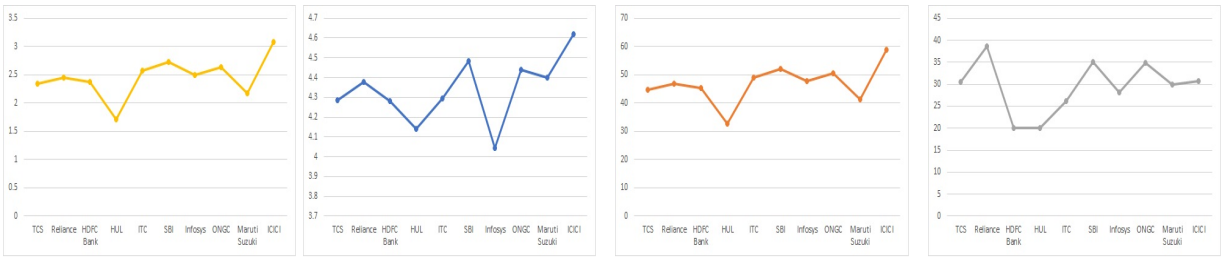
[1] Most of the data analysis, interpretation and modelling is done by using Python 3.6.4 and Microsoft Office Excel 2016.
[2] Source - Official website of National Stock Exchange of India.

TABLE 1: Entropy and Volatility results for selected Indian Companies

	TCS	RELIANCE	HDFC BANK	HUL	ITC	SBI	INFOSYS	ONGC	MARUTI SUZUKI	ICICI
Daily Historical Volatility	2.34	2.45	2.38	1.71	2.57	2.72	2.49	2.64	2.17	3.08
Annualized Historical Volatility	44.73931994	46.73316927	45.44156973	32.64344619	49.11858614	51.9632587	47.6258941	50.43246184	41.36674589	58.8840549
Implied Volatility	30.5938721	38.57421875	20.01953125	20.11108398	26.18408203	35.1257324	28.0456543	34.79003906	30.01403809	30.76171875
Entropy	4.285956147	4.379801726	4.280798401	4.142849766	4.295392422	4.483919655	4.042877496	4.43936491	4.401005022	4.619018274

TABLE 2: Correlation Matrix for Indian Companies

	Entropy	Annualized Historical Volatility	Implied Volatility
Entropy	1		
Annualized Historical Volatility	0.65639891	1	
Implied Volatility	0.547485879	0.5137703	1



(a) Daily Historical Volatility (b) Entropy (c) Annualized Historical Volatility (d) Implied Volatility

FIGURE 1: Comparative analysis of the entropy and volatility for Indian Companies

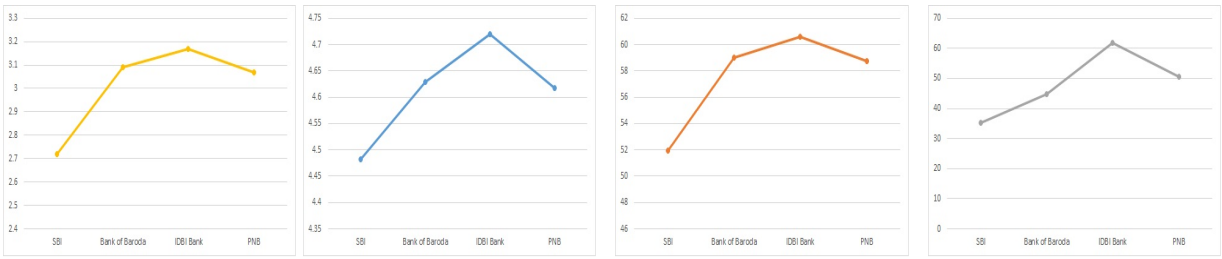
TABLE 3: Public Sector Banks

Entropy and Volatility results

	SBI	Bank of Baroda	IDBI Bank	PNB
Daily Historical Volatility	2.72	3.09	3.17	3.07
Annualized Historical Volatility	51.9632587	59.01722391	60.61496826	58.71762192
Implied Volatility	35.1257324	44.79980469	61.76757813	50.65917969
Entropy	4.482365622	4.628906387	4.720556759	4.616763381

Correlation Matrix

	Entropy	Annualized Historical Volatility	Implied Volatility
Entropy	1		
Annualized Historical Volatility	0.962957291	1	
Implied Volatility	0.951469855	0.877646967	1



(a) Daily Historical Volatility (b) Entropy (c) Annualized Historical Volatility (d) Implied Volatility

FIGURE 2: Comparative analysis of the entropy and volatility for Public Sector Banks

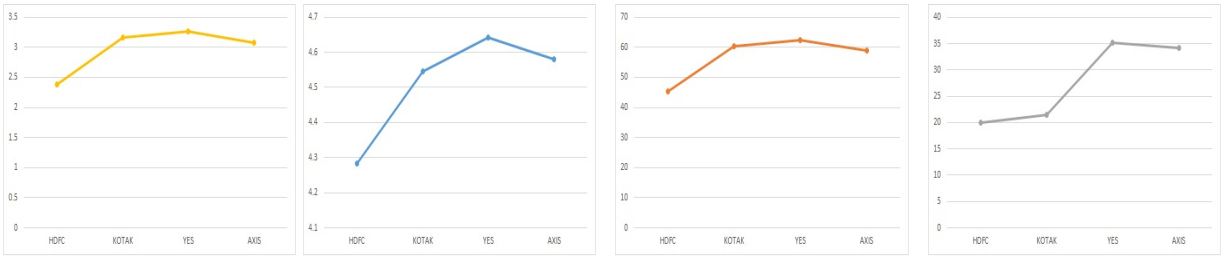
TABLE 4: Private Sector Banks

Entropy and Volatility results

	HDFC	KOTAK	YES	AXIS
Daily Historical Volatility	2.38	3.16	3.26	3.08
Annualized Historical Volatility	45.44156973	60.40155195	62.29837139	58.92868316
Implied Volatility	20.01953125	21.45385742	35.15625	34.11865234
Entropy	4.282262489	4.546316805	4.641434414	4.581236584

Correlation Matrix

	Entropy	Annualized Historical Volatility	Implied Volatility
Entropy	1		
Annualized Historical Volatility	0.983360564	1	
Implied Volatility	0.771749481	0.643492335	1



(a) Daily Historical Volatility (b) Entropy (c) Annualized Historical Volatility (d) Implied Volatility

FIGURE 3: Comparative analysis of the entropy and volatility for Private Sector Banks

Conclusion

The main idea of this paper is that entropy can be used as an ure for volatility. The decisive advantage of this approach resides in its ability to capture disorder or uncertainty in the system without putting any constraint on the probability distribution function. We have also modelled implied volatility as a linear combination of historical volatility and entropy, and found that the model was heavily dependent on the entropy.

REFERENCES

[1] S. Bentes, R. Menezes, Entropy: A new measure of stock market volatility?, J. Phys.: Conf. Ser. 394 (1) (2012) 12-33. doi:10.1088/1742-6596/394/1/012033.

[2] R. Shiller, Do Stock Prices Move Too Much to be Justified by Subsequent Changes in Dividends?, The American Economic Review 71 (3) (1981) 421-436. doi:10.3386/w0456.

[3] G. Schwert, Why Does Stock Market Volatility Change Over Time?, The Journal of nance 44 (5) (1989) 1115-1153. doi:10.1111/j.1540-6261.1989.tb02647.x.

[4] J. C.Hull, Options, Futures and other Derivatives 8th edition, Pearson Education,Inc., USA, 2012.

[5] F. Black, M. Scholes, The pricing of options and corporate liabilities, J.Political Econ. 81 (3) (1973) 637-659. doi:10.1086/260062.

[6] C. E. Shannon, A mathematical theory of communication, Bell System Technical Journal 27 (3) (1948) 379-423. doi:10.1002/j.1538-7305.1948.tb01338.x.

[7] L. Gulko, The Entropy theory of stock option pricing, International Journal of Theoretical and Applied Finance 2 (3) (1999) 331-355. doi:10.1142/S0219024999000182.

Fuzzy Logic Modelling of Degradation of Cement Mortar in Aggressive Environment

Kavita Verma, Shobha Ram, Alok Verma

Abstract*: Degradation of cement mortar takes place in the acidic aggressive environment due to the alkaline nature of hydration compounds of cement. This happens to depend on many factors such as the aggressiveness of environment, time of exposure, mortar mix composition, etc. A study was carried out to appreciate the mechanism of such degradation and an effort has been made to classify various patterns in the degradation process. Due to roles of various factors in such a process, a definite nature and degree of fuzziness creeps in this mechanism. It has been the motive of this study to accept this fact and apply fuzzy logic concepts to such a mechanism comprising of various stages of degradation in the form of different effects which could be visually observed. In this study, it is intended to define degradation stages in cement mortar more rationally by the concept of fuzzy logic and to find out the optimum number of categories in which various types of effects should be divided so that various degradation effects being produced in the material are rightly judged and calibrated with the correct 'structural' condition of the material.

Keywords: Mortar, Sulfate Attack, Visual Rating, Fuzzy Logic

I. INTRODUCTION

Visual observations of samples, being tested in laboratories are made to understand the damage in cement-based construction materials such as cement mortar etc., to understand the damage mechanisms related to the deterioration of such materials in aggressive environments. Such observations also provide us with the progress of deterioration and sequence of the happening of various types of defects in such materials. Various types of such damage sequences in the form of visual ratings have been provided in cement mortar and concrete (Irassar et al. 2009; Zeng at al 2018; Skaropoulou et al. 2009; Tamimi et al 2003; Brown et al 2004; Al-Amoudi 2002; Michael Thomas 2008; Abdelmseeh 2008). Though visual observations are helpful in making us understand the process of damage in cement-based materials, their results are many times of a fuzzy nature. Due to this, it becomes somewhat difficult to provide clear conclusions from the results.

Revised Manuscript Received on September 15, 2019

Kavita Verma, Civil Engineering Department, Gautam Buddha University, Greater Noida, Uttar Pradesh (201308), India. Email: kavitaverma.22k@gmail.com

Dr. Shobha Ram, Civil Engineering Department, Gautam Buddha University, Greater Noida, Uttar Pradesh (201308), India. Email: shobharam@gbu.ac.in

Prof. Alok Verma, Civil Engineering Department, Delhi Technological University, Delhi (110042), India. Email: alokverma_dce@hotmail.com

In the case of concrete, many effects have been seen in various stages of deterioration. These have been defined as subsurface cracks (Stock et al 2002), fine cracks (Planel et al 2006), spalling of surface material (Hartshorn et al 1999; Al-Akhras 2006; Zhou et al 2006), blistering of surface (Hartshorn et al 1999; Collett et al 2004), crazing (Al-Dulaijan et al 2003), loosening of material (Zhou et al 2006) etc. It has been appreciated that length expansion takes place in a specimen which may lead to the development of stresses (Hartshorn et al 1999). Sometimes, surface cracks are combined with other effects (Al-Akhras 2006). Effects such as swelling of corners, a complete breakdown of samples, appearance of soft pulpy mass of mortar, debonding of matrix from aggregate particles, extreme distress, onion peeling type degradation of samples have been seen by many researchers in the past (Gallop and Taylor 1996; Planel et al 2006; Hooton 2008; Zhou et al 2006; Al-Akhras 2006; Lee et al 2008). Many other effects such as erosion of faces, change in colour, loss of cohesion, softening of an external layer of mortar specimen have also been reported (Zhou et al 2006; Hartshorn et al 1999; Collett et al 2004; Planel et al 2006; Torres et al 2004; Haynes 2008).

All cements, including the sulfate-resisting cement are vulnerable to sulfate attack (Rasheeduzzafar 1992, Al-Amoudi 1998). Rate and intensity of attack depend on cement type, sulfate type & concentration, quality of material, exposure conditions, etc. (Al-Amoudi 2002). Cement mortar also is subjected to the same types of effects, as in cement concrete, due to cement being the same strength imparting material in both the cases. Many researchers have provided the occurrence of various effects in cement mortar samples with the help of visual observations. (Skaropoulou et al. 2009) have considered the long-term behavior of Portland limestone cement mortars exposed to magnesium sulfate attack. The study has considered various stages of degradation and these stages have been defined in the form of various effects appreciated by visual observations. Nine different effects have been shown and it has been observed by the data of visual observations that these effects take place in the same sequence in various types of cement mortar mixes in the 1.8%MgSO4 solution. It has also been observed that the complete deterioration of mortar mixes takes place at approximately the same time, except only one mix of mortar. Effect of the aggressive environment has been reported in the context of Ultrasonic pulse velocity test results and mass loss taking place in aggressive environment.

An average change of ultrasonic pulse velocity values to the extent of 15 to 19% has been shown to have taken place in a time span of 6 to 12 months. Mass loss up to an extent of more than 50% is also reported (Skaropoulou et al. 2009).

Effect of Magnesium sulphate solution on sulfate resistance of geopolymer mortar has been investigated (Elyamany et al, 2018) and the deteriorating visual effects have been correlated with parameters such as weight change, compressive strength, expansion strain, etc. The combined effect of chloride and sulfates has been investigated by (Abdalkader et al 2015) and visual observations have been reported. It has been observed that the effects of deteriorating aggressive environment are first observed in the rounding of edges and corners. Change of mass, color changes and spalling of mass from mortar specimens have also been reported. It is seen that the effects reported by visual observations in the most of the cases are broadly in a uniform sequence. (Al-Dulaijan et al. 2003) have investigated the effect of sulfate environments of various concentrations on cement mortar specimen and have recorded visual observations. Very high concentrations of sulfates, occurring in saline soils, have been considered in the study and strength reduction has been correlated to the visual effects. (Chen et al. 2017) have considered the deteriorating mechanism of plain and blended cement mortars exposed to sulfate attack and has mentioned various effects such as softening, expansion, cracking, spalling and disintegration of cement-based materials. Visual observations in the case of cement mortar have been reported (Hartshorn et al 1999) and similar types of effects have been seen to be happening in all the degrading types of environments. Visual effects on cement mortar specimens have been reported in various types of aggressive environments (Irassar et al 2009, Aye et al. 2011, Al-Amoudi,2002, Kwasny et al.2018, Maes et al 2017, Pastor et al.2018, Planel et al 2006, Zhou et al. 2006, Stock et al 2002). A summary of such details is provided in Table 1 below.

Table- I: Summary of visual observations in cement mortar

Researcher's Name	**Curing Environment**	**Effects recoded after visual examination**
Hartshorn et al 1999	*Na_2SO_4 solution*	*No visible deterioration, White powdery coating, crazing of surface, Expansion, cracking and spalling, beginning to spall, extensive spalling, Surface softening, no cracking but some corrosion at edges, Blisters containing grey white mush, beginning to crack, Partial disintegration Cracking and spalling, Disintegration Spalling and softening.*
Al-Amoudi, 2002	*$MgSO_4$ & Na_2SO_4 solution*	*Eating away of the hydrated cement paste and progressively reducing it to cohesionless granular mass leaving the aggregates exposed – reduction of cross-sectional area –decrease in strength, Expansion and cracking, Onion peeling type of scaling or shelling of the surface in successive layers in the form of delamination and cracking*
Stock et al., 2002	*10000 ppm Na_2SO_4 solution*	*No damage, Extreme damage*
Al-Dulaijan et al., 2003	*Sulfate conc. of 1 to 4%*	*No damage, Cracking, minor spalling, spalling of the surface skin at the corners, intense spalling*
Zhou et al., 2006	*1.6 g SO_4 /L*	*Formation of a white mushy deposit on the top surface of the cubes, Blistering, spalling at the corners and edges of the cubes. aggregate exposure, color change, very fine crystals were observed on the surfaces*
Planel et al., 2006	*15 mmol/L of Na_2SO_4 solution*	*Cracking, surface delamination, leached surfaces, surface spalling*
Skaropoulou et al., 2009	*1.8% $MgSO_4$*	*No visible deterioration, some deterioration at corners, deterioration at corners and some cracking along the edges, deterioration at corners and cracking along the edges, cracking and expansion, bulge of surfaces, extensive cracking, extensive spalling, and complete damage.*
Irassar et al., 2009	*Na_2SO_4 & MgSO4 solution*	*No visible deterioration, Deterioration at corners and edges, Cracking along the edges, Extensive cracking and expansion, Spalling and disintegration of surfaces*
Seung-Tae Lee 2009	*$MgSO_4$ Solution*	*Surface damage at the corners of the mortar, white powdery material deposited on the faces of mortars, spalling and cracking at the corners, edges and faces, small spalling only at the corners and edges, some surface damages (i.e., spalling and delamination)*

Aye et al., 2011	*$MgSO_4$ & 10% Na_2SO_4 solution*	*No visual damage, loss of cohesionless particles from the surface of the specimens, no other damages (such as cracking and spalling), weight loss, extensive cracking, efflorescence, loss of cohesionless particles*
Maes et al., 2014	*$MgSO_4$ solution*	*Spalling of the outermost layers, some small cracks, extensive cracking, crumbling., mass changes, spalling of surface layer*
Chen et al., 2017	*10% Na_2SO_4 solution*	*No damage, Salt efflorescence, surface scaling and flaking, cracking, weight loss*
Q. Zeng et al., 2018	*50 g/L Na_2SO_4 and 44.2 g/L $MgSO_4$ Solution*	*No visible deterioration, Some deterioration at corners, Deterioration at corners and the edges, Deterioration at corners and some cracking along the edges, Cracking and expansion along the edges, Serious cracking and expansion, Spalling and expansion of surface, Grey pulp at corners, edges and surface, Amount of grey pulp at specimens, Complete damage*
Elyamany et al., 2018	*10% $MgSO_4$*	*Spalling of edges, change in shape, micro cracks and pores, visible cracks, soft and powdery white deposits on surface which became harder with time, surface erosion, weight loss*
Kwasny et al., 2018	*H_2SO_4 & HCl solutions*	*No sign of discoloration, expansion or cracking on the surface, microcracks, curled, broken, and longitudinal, layer of white precipitates*
Pastor et al., 2018	*15% Na_2SO_4 solution*	*Increase in roughness, crystallization of salts, loss of small pieces of material, and cracks*

II. PROGRAMME OF THE STUDY

It is understood that visual observations of effects in durability studies portray the strength conditions of a material. That is why, visual effects are recorded in such types of studies. In the case of cement concrete, many researchers have specified different visual ratings based on which other characteristics of concrete have been calibrated. It may be appreciated that a particular type of visual effect denotes a range of a particular condition and in that range the 'degree' of that effect is to be further defined based on clarity of information about tests and experience, expertise etc., of the researcher. For example, the effect of color change may be defined as small, medium or a large color change. Boundaries between these sub-effects become fuzzy as many times it may be difficult to put a particular sample in the correct category. Many times, it may be easier to recognize the effect but it may be more difficult to decide the intensity of that effect. In this study, it is intended to find out the optimum number of categories in which various types of effects should be divided so that various effects being produced in the material are rightly judged and calibrated with the correct 'structural' condition of the material.

The present study considered the durability conditions of 1:4 and 1:6 mix cement mortar cubes of size 10 cm side in H2SO4 solution of normality N1 and N2. Fig.1 shows some samples of this study showing different degrees of various effects occurring in samples. In some samples, simultaneous occurrence of various effects can also be recognized. If color change is taken into consideration, it may not be definitely said that all mortar specimens are at the same level of deterioration. Many times, it may not be easy to compare two samples based on a particular effect. In many cases, two or more effects are there in a sample and the decision making about existing degradation may be difficult to be taken. This fact impels for a fuzzy based approach to be taken up.



Fig.1. Cement mortar cubes in different stages of degradation

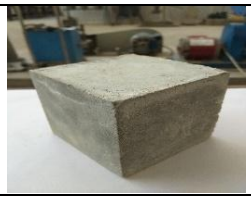


Fig. 2. No visible deterioration



Fig. 3. Efflorescence: low, medium & high



Fig. 4. Color change: low, medium & high



Fig. 5. Falling of edges: low, medium & high



Fig. 6. Falling of corners: low, medium & high



Fig. 7. Surface Roughness: low, medium & high



Fig. 8. Blistering of surface: low, medium & high

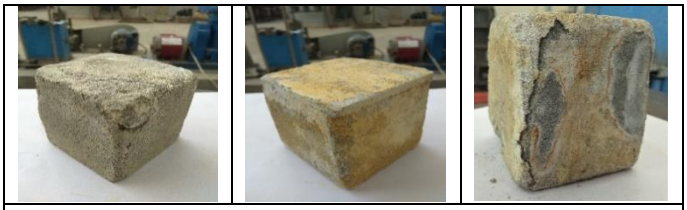


Fig. 9. Cracks: low, medium & high



Fig. 10. Spalling of mass: low, medium & high



Fig. 11. Softening of cubes: low, medium & high



Fig. 12. Complete Damage: low, medium & high

The visual examination of the samples performed on cement mortar at regular intervals of time. Some significant modifications, such as formation of white mushy layer deposited on the surface of the mortar i.e., efflorescence, changes in surface color and texture, formation of coatings, falling of edges and corner, surface roughness**,** blistering of surface, cracking and spalling of mass, softening of cubes and complete deterioration were recorded and are shown in Fig. 3 to Fig. 12 in the form of low, medium and high ratings. Fig. 2 displays a cement mortar sample which is having no signs of visible deterioration. Though such a case may not require any need of fuzzy logic to appreciate a condition of no deterioration, this case is taken here for inclusion of all cases for the sake of complete treatment. To assess the deterioration performance of the cement-based structures, many researches have included a description of the visual appearance of degraded specimens in the form of visual ratings (Irassar 2009) and some of these are presented in Table I. Visual ratings made on the basis of these reported effects should have an optimum number of classes to contain all effects being produced due to degradation. If a visual rating has a large number of classes, many simultaneously occurring effects may not be considered. If only a few classes are provided in the visual rating scheme, effects may not be graded accurately to define degradation of material.

In this study based on fuzzy logic, output confidence in the form of likely accuracy in recoding the actual degradation in a sample is calculated. Mamdani type inference system in MATLAB was used in this study. Two input parameters – degradation classes and site & exposure conditions – are considered. Membership functions for degradation classes are defined in terms of Gaussian functions while those for site & exposure conditions are defined as triangular functions. Three cases of input 1 – degradation classes with 3, 6 and 9 categories – have been considered. Recognition of site & exposure conditions may require human attributes such as expertise, experience and qualification related to deterioration aspects of cement-based materials in aggressive environments. Three categories – poor, average and good – have been prescribed to classify this input. Input 2, based on exposure condition, consists 3 membership function that remain same for all cases. The output has been considered in terms of confidence which can be placed in a situation based on degradation classes and site & exposure conditions. It also is defined in terms of gauss membership functions. Values of 0 and 1 for a particular input / output parameter may respectively show a ‘complete uncertain situation’ and ‘accurately defined situation’.

Output confidence for low-low, low-high, high-low and high-high values of input 1-input 2 values have been determined. Rule view for a test case when the values of confidence in input 1 and input 2 are varied systematically is shown in Fig. 17. Fuzzy inference systems, used in this study, is given below.

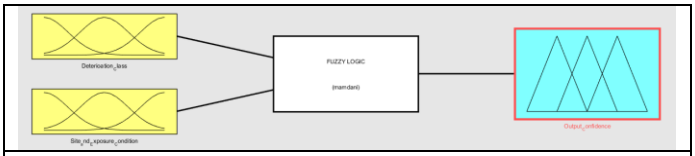


Fig. 13. Fuzzy system with 2 inputs, 1output and 27 rules

III. FUZZY INFERENCE SYSTEM

Name='fuzzylogic'
Type='mamdani'
Version=2.0
NumInputs=2
NumOutputs=1
NumRules=27
AndMethod='min'
OrMethod='max'
ImpMethod='min'
AggMethod='max'
DefuzzMethod='centroid'

[Input1]
Name='deterioration class'
Range=[0 1]
NumMFs=9
MF1='stage1':'gaussmf',[0.04946 0.00576]
MF2='2':'gaussmf',[0.03015 0.1495]
MF3='3':'gaussmf',[0.03075 0.2493]
MF4='4':'gaussmf',[0.02722 0.3645]
MF5='5':'gaussmf',[0.02765 0.4665]
MF6='6':'gaussmf',[0.0272 0.57]
MF7='7':'gaussmf',[0.02913 0.6612]
MF8='8':'gaussmf',[0.0288 0.769]
MF9='9':'gaussmf',[0.09343 1]
[Input2]
Name='site & exposure condition'
Range=[0 1]
NumMFs=3
MF1='poor':'trimf',[-0.4929 -0.0071 0.3651]
MF2='average':'trimf',[0.1576 0.4994 0.8307]
MF3='good':'trimf',[0.634 1.01 1.5]
[Output1]
Name='output_confidence'
Range=[0 1]
NumMFs=3
MF1='poor':'gaussmf',[0.15 0]
MF2='average':'gaussmf',[0.15 0.5]
MF3='good':'gaussmf',[0.15 1]

[Rules]

1 1, 1 (1) : 1 1 2, 2 (1) : 1 1 3, 3 (1) : 1 2 1, 1 (1) : 1 2 2, 2 (1) : 1 2 3, 3 (1) : 1 3 1, 1 (1) : 1 3 2, 2 (1) : 1 3 3, 3 (1) : 1 4 1, 1 (1) : 1 4 2, 2 (1) : 1 4 3, 3 (1) : 1 5 1, 1 (1) : 1 5 2, 2 (1) : 1 5 3, 3 (1) : 1 6 1, 1 (1) : 1 6 2, 2 (1) : 1 6 3, 3 (1) : 1 7 1, 1 (1) : 1 7 2, 2 (1) : 1 7 3, 3 (1) : 1 8 1, 1 (1) : 1 8 2, 2 (1) : 1 8 3, 3 (1) : 1 9 1, 1 (1) : 1 9 2, 2 (1) : 1 9 3, 3 (1) : 1

IV. MEMBERSHIP FUNCTIONS

Membership functions showing different categories of the input 1 parameter of 3, 6 and 9 visual classes are shown in Fig. 14 to Fig. 16.

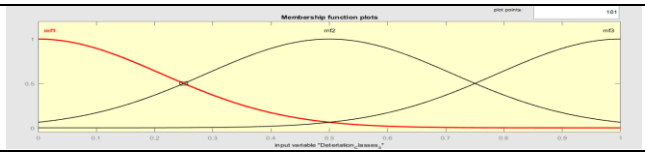


Fig. 14. Membership function for Input1 (3 classes)

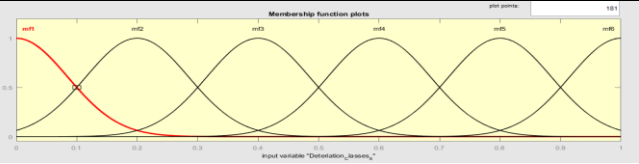


Fig. 15. Membership function for Input1 (3 classes)

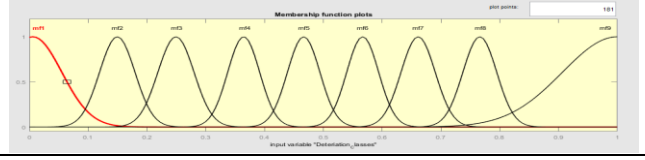


Fig. 16. Membership function for Input1 (9 classes)

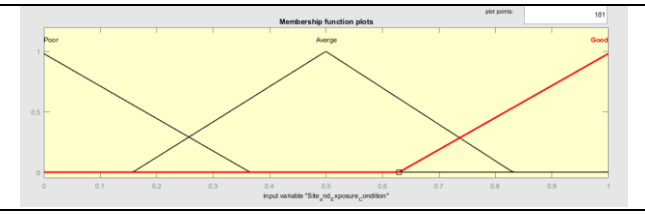


Fig. 17. Membership function for Input2 (3 classes)

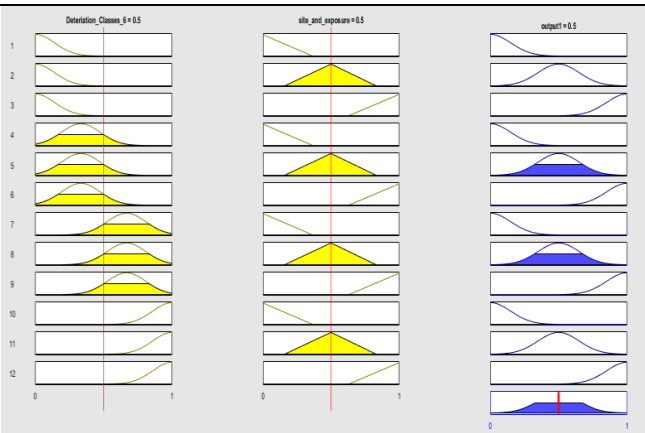


Fig. 18. Rule view for particular values of input and output

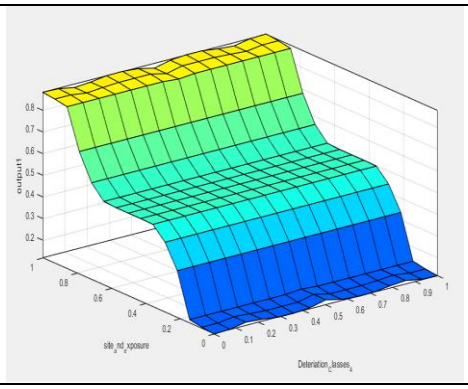


Fig. 19. Surface view

V. RESULT AND DISCUSSION

A particular result, obtained from rule view of MATLAB, is shown in Fig. 17. Surface view for input and output is shown in Fig. 18. In such a diagram, only two input parameters may be considered with the output at a time. From the rule view and surface view it is observed the confidence of output depends, to a great extent, on input 2. This type of variation has been seen in all three cases of same type. It is clearly observed that input 2 affects the confidence in output to a greater extent than input 1. It means output confidence or 'accuracy of decision for characterizing depends on exposure conditions, the availability of information with respect to site & material more than number of categories in a visual rating system. Output confidence with respect to a systematic change of input1 and input 2 has been shown in Table II.

From the data given in Table II, following conclusions may be noted:

Medium output confidence in all the three cases is obtained when value close to 0.5 is given to both input 1 and input 2, as shown in row is 1st row. Output confidence value does not change in all the cases when input 1 has been taken to be very low while input 2 remains as such as, shown in 2nd row. Value of output confidence does not change in all the cases when value of input 1 has been taken to be high while input 2 remains as such, as shown in 3rd row. Value of output confidence lowers down tremendously in all the cases when value of input 1 has been restored as its initial value while input 2 has been assigned a low value, as shown in 4th row. Value of output confidence rises up in all the cases when value of input 1 has been taken as its initial value, while input 2 has been assigned a high value as shown in 5th row.

Output confidence in the results of visual examination in field studies and laboratory, depends on the confidence generated due to the availability of accurate information rather than number of classes in a visual rating system. The meaning of a particular class becomes vague and also the reach of classes becomes very broad, if the number of classes is too low. It may be difficult to accurately select the right deterioration stage for the particular material if visual rating classification system consists of too many classes. In such case, boundaries between adjoining classes become more and more fuzzy. This problem may be understood from the concepts given by (Klir and Yuan 2001), 'opted uncertainty' and 'forced uncertainty' used in fuzzy systems analysis. In the visual rating system, if the number of classes is reduced by merging, it becomes a case of 'opted uncertainty'. For example, if only two stages are there in a visual rating classification system then deterioration may be given either as 'nil' or 'full'. Having too many classes in visual rating also gives an idea of 'Forced uncertainty'. For example, a rating system having too many classes may be dependent on the existence of a crack. In that condition, presence of very minute cracks which may not be observed with the help of available instruments may not help us to designate the appropriate deterioration rating corresponding to that class of cracking.

Some common effects at various stages, given in table I, may be combined and suggested, as given in Table III. The boundaries of the categories still remain fuzzy and these overlap one another by some degree. That is why common effects may be indicated in different classes. It may be seen that color changes appear in stage 2, 3 and 4 but the degree of color change may be different. Likewise, disintegration of samples in different stages comes with different adjectives – partial or complete disintegration. It is suggested that a degradation stage may be allocated to a sample based on all likely visual effects as per that stage in question. All visual effects in a combined way, may indicate the presence of a stage of degradation which, in turn, may point towards a 'structural condition'.

Table- II: Output with variations in input

S. No.	Rating system with 3 classes			Rating system with 6 classes			Rating system with 9 classes		
	Number of classes	Exposure	Result confidence	Number of classes	Exposure	Result confidence	Number of classes	Exposure	Result confidence
1.	0.501	0.512	0.5	0.501	0.512	0.5	0.501	0.512	0.5
2.	0.012	0.512	0.5	0.012	0.512	0.5	0.012	0.512	0.5
3.	0.989	0.512	0.5	0.989	0.512	0.5	0.989	0.512	0.5
4.	0.501	0.001	0.117	0.501	0.001	0.14	0.501	0.001	0.143
5.	0.501	0.998	0.883	0.501	0.998	0.86	0.501	0.998	0.857

Some common effects at various stages, given in table 1, may be combined and suggested, as given in Table 3. The boundaries of the categories still remain fuzzy and these overlap one another by some degree. That is why common effects may be indicated in different classes. It may be seen that colour changes appear in stage 2, 3 and 4 but the degree of color change may be different. Likewise, disintegration of samples in different stages comes with different adjectives – partial or complete disintegration. It is suggested that a degradation stage may be allocated to a sample based on all likely visual effects as per that stage in question. All visual effects in a combined way, may indicate the presence of a stage of degradation which, in turn, may point towards a 'structural condition'.

Table-III: Suggested visual rating system for degradation of cement mortar in aggressive environment

Stage	Likely visual effects
1	No visible deterioration, efflorescence, hairline cracks, roughening of surface
2	Cracking due to expansion in material, initial color changes from grey color of cement, spalling of edges & corners of samples
3	Partial disintegration, surface peeling off, extensive softening and spalling, color changes
4	Intense color changes, extensive softening and spalling, reduction of cross section, considerable to complete disintegration

VI. CONCLUSIONS

Following conclusions may be drawn from this study.

1. A visual rating system should be accurate as well as efficient in deciding the degradation stage of a sample. Accordingly, it should contain an optimum number of classes or stages. In this study four classes are found to be necessary and sufficient for the purpose.
2. Every different effect representing degradation should be clearly defined in a visual rating system. Due to fuzzy nature of conditions representing adjoining stages, the use of fuzzy logic is beneficial in such cases of decision making.
3. In deciding deterioration stage for a sample, it should be ensured that the sample conforms broadly to given possible visual effects of a stage.

ACKNOWLEDGMENT

The author wishes to express his sincere thanks to several researchers for reporting their results appeared in various publications.

REFERENCES

1. A.K. Al. Tamimi and M. Sonebi, "Assessment of Self-compacting concrete immersed in acidic solutions," Journal of Materials in Civil Engineering, ASCE, Volume 15, Issue 4, August, 2003, pp. 354-357.
2. Omar Saeed Baghabra Al-Amoudi, "Sulfate attack and reinforcement corrosion in plain and blended cements exposed to sulfate environments," Building and Environment, Volume 33, Issue 1, 1998, pp. 53–61.
3. S.U. Al-Dulaijan, M. Maslehuddin, M.M. Al-Zahrani, A.M. Sharif, M. Shameem and M. Ibrahim, "Sulfate resistance of plain and blended cements exposed to varying concentrations of sodium sulfate," Cement & Concrete Composites, Volume 25, Issues 4–5, 2003, pp. 429–437.
4. Ashraf H.M. Abdalkader, Cyril J. Lynsdale, and John C. Cripps, "The effect of chloride on cement mortar subjected to sulfate exposure at low temperature," Construction and Building Materials, Volume 78, 2015, pp. 102-111.
5. G. Collett, N.J. Crammond, R.N. Swamy, and J.H. Sharp, "The role of carbon dioxide in the formation of thaumasite," Cement and Concrete Research, Volume 34, Issue 9, 2004, pp. 1599-1612.
6. E.F. Irassar, "Sulfate attack on cementitious materials containing limestone filler — A review," Cement and Concrete Research, Volume 39, Issue 3, 2009, pp. 241–254.
7. Fei Chen, Jianming Gao, Bing Qi and Daman Shen, "Deterioration mechanism of plain and blended cement mortars partially exposed to sulfate attack," Construction and Building Materials, Volume 154, 2017, Pages 849-856.
8. R.S. Gollop and H.F.W. Taylor, "Microstructural and Microanalytical Studies of Sulfate Attack. V. Comparison of Different Slag Blends," Cement and Concrete Research, Volume 26, Issue 7, 1996, pp. 1029-1044.
9. Hafez E. Elyamany, Abd Elmoaty, M. Abd Elmoaty and Ahmed M. Elshaboury, "Magnesium sulfate resistance of geopolymer mortar," Construction and Building Materials, Volume 184, 2018, pp. 111-127.
10. S.A. Hartshorn, J.H. Sharp and R.N. Swamy, "Thaumasite formation in Portland-limestone cement pastes," Cement and Concrete Research, Volume 29, Issue 8, 1999, Pages 1331-1340.
11. Harvey Haynes, Robert O'Neill, Michael Neff and P. Kumar Mehta, "Salt weathering distress on concrete exposed to sodium sulfate environment," ACI Material, Volume 105, Issue 1, 2008, pp. 35-48.
12. R.D. Hooton, "Bridging the gap between research and standards," Cement and Concrete Research," Volume 38, Issue 2, 2008, pp. 247-258.
13. Jose L. Pastor, J. Marcos Ortega, Miguel A. Climent, and Isidro Sánchez, "Skin friction coefficient change on cement grouts for micropiles due to sulfate attack," Construction and Building Materials, Volume 163, 2018, pp. 80-86.
14. Jacek Kwasny, Timothy A. Aiken, Marios N. Soutsos, John A. McIntosh and David J. Cleland, "Sulfate and acid resistance of lithomarge-based geopolymer mortars," Construction and Building Materials, Volume 166, 2018, pp. 537-553.
15. Mathias Maes and Nele De Belie, "Resistance of concrete and mortar against combined attack of chloride and sodium sulphate,"Cement and Concrete Composites," Volume 53, 2014, pp. 59-72.
16. Mathias Maes and Nele De Belie, "Influence of chlorides on magnesium sulphate attack for mortars with Portland cement and slag-based binders," Construction and Building Materials, Volume 155, 2017, pp. 630-642.
17. Jan Małolepszy and Ewelina Grabowska, "Sulphate Attack Resistance of Cement with Zeolite Additive," Procedia Engineering, Volume 108, 2015, pp. 170-176.
18. Michael Thomas, Kevin Folliard, Thano Drimalas and Terry Ramlochan, "Diagnosing delayed ettringite formation in concrete structures," Cement and Concrete Research, Volume 38, Issue 6, 2008, pp. 841-847.
19. Nabil M. Al-Akhras, "Durability of metakaolin concrete to sulfate attack," Cement and Concrete Research, Volume 36, Issue 9, 2006, pp. 1727-1734.
20. Omar S. Baghabra Al-Amoudi, "Attack on plain and blended cements exposed to aggressive sulfate environments," Cement & Concrete Composites, Volume 24, Issue 3–4, 2002, pp. 305-316.
21. Paul Brown, R.D. Hooton and Boyd Clark, "Microstructural changes in concretes with sulfate exposure," Cement and Concrete Composites, Volume 26, Issue 8, 2004, pp. 993-999.
22. D. Planel, J. Sercombe, P. Le Bescop, F. Adenot and J.-M. Torrenti, "Long-term performance of cement paste during combined calcium leaching–sulfate attack: kinetics and size effect," Cement and Concrete Research, Volume 36, Issue 1, 2006, pp. 137-143.
23. Rasheeduzzafar, "Influence of cement composition on concrete durability," ACI Material Journal, Volume 89, Issue 6, 1992, pp. 574–86.
24. S. Thokchom, P. Ghosh and S. Ghosh, "Performance of fly ash based geopolymer mortars in sulphate solution," Journal of Engineering Science and Technology Review, Volume 3, Issue 1, 2010, pp. 36–40.
25. Seung-Tae Lee, "Influence of recycled fine aggregates on the resistance of mortars to magnesium sulfate attack," Waste Management, Volume 29, Issue 8, 2009, pp. 2385-2391.

26. A. Skaropoulou, S. Tsivilis, G. Kakali, J.H. Sharp and R.N. Swamy, "Long term behavior of Portland limestone cement mortars exposed to magnesium sulfate attack," Cement and Concrete Composites, Volume 31, Issue 9, 2009, pp. 628-636.
27. S.R Stock, N.K Naik, A.P Wilkinson and K.E Kurtis, "X-ray microtomography (microCT) of the progression of sulfate attack of cement paste," Cement and Concrete Research, Volume 32, Issue 10, 2002, pp. 1673-1675.
28. T. Bakharev, "Durability of geopolymer materials in sodium and magnesium sulfate solutions," Cement and Concrete Research, Volume 35, Issue 6, 2005, pp. 1233-1246.
29. Thidar Aye and Chiaki T. Oguchi, "Resistance of plain and blended cement mortars exposed to severe sulfate attacks," Construction and Building Materials, Volume 25, Issue 6, 2011, pp. 2988-2996.
30. S.M. Torres, C.A. Kirk, C.J. Lynsdale, R.N. Swamy and J.H. Sharp, "Thaumasite–ettringite solid solutions in degraded mortars," Cement and Concrete Research, Volume 34, Issue 8, 2004, pp. 1297-1305.
31. V. Assaad Abdelmseeh, J. Jofriet and G. Hayward, "Sulphate and sulphide corrosion in livestock buildings, Part I: Concrete deterioration," Biosystems Engineering, Volume 99, Issue 3, 2008, pp. 372-381.
32. A. Verma, M. Shukla and A. K. Sahu, "Effect of number of classes in a visual rating for sulphate attack," International Journal of Civil Engineering and Technology, Volume 4, Issue 1, 2013, pp. 165-181.
33. Qianpu Zeng, Chong Wang, Yaoling Luo, Chao Yu, Qian Huang and Chaoqun Luo, "Effect of temperatures on TSA in cement mortars under electrical field," Construction and Building Materials, Volume 162, 2018, pp. 88-95.
34. Zhou, J. Hill, E.A. Byars, J.C. Cripps, C.J. Lynsdale and J.H. Sharp, "The role of pH in thaumasite sulfate attack," Cement and Concrete Research, Volume 36, Issue 1, 2006, pp. 160-170.

AUTHORS PROFILE



Kavita Verma is a Ph.D. student in civil engineering at Gautam Buddha University. Her research interests in the area of durability of cement mortar in aggressive environment and use of fuzzy logic and neural networking in the area of durability studies.



Shobha Ram is an Assistant Professor and Head of department in Civil Engineering at Gautam Buddha University, Greater Noida. His area of interest is Hydrology, Environmental Engineering, Water Pollution Control, Water Resource Management & Water quality monitoring, Air pollution, Structural materials.



Alok Verma is a Professor in Civil Engineering at Delhi Technological University, Delhi (India). His research interests include Earthquake technology and use of fuzzy logic and neural networking in the area of durability studies. He has been an author for a review book on Civil Engineering and a co-author for many reports and proceedings. He is a member of the committee CED – 39 of Bureau of Indian Standards, New Delhi (India) considering revision of Indian standard guidelines and code provisions in Earthquake Engineering.

2019 IEEE Intelligent Transportation Systems Conference (ITSC)
Auckland, NZ, October 27-30, 2019

GPS Based Adaptive Traffic Light Timings and Lane Scheduling

Krishna Kant Sharma[1] and Indu S[2]

Abstract— **In today's era, it has become easier to collect traffic data using GPS enabled devices. In cities, a good proportion of people use Map Services like Google Maps in their daily commute. Google's Geocoding and Direction APIs can easily track users' geolocation and the route they are going to take in their commute. The scope of this work is to develop and create a mobile application which aids in the collection of GPS data from the users and use it to allocate fair green light timings, schedule proper lane departures and suggest lanes accordingly to the users approaching an intersection. Assuming that the drivers will follow the suggestion of the mobile application, this will reduce congestion and delays on roads with high traffic flow by prioritizing and assigning them the maximum green light time. The lanes with little traffic are assigned less green light time while the ones with no traffic are skipped for the next cycle. In order to validate our work, we used AnyLogic [9] to simulate different traffic scenarios. We found that our method allocates fair green light timings and lane scheduling, thus significantly reducing the traffic congestion on the roads with higher traffic flow. The main motivation behind the work is to develop a system which is cheap to implement and can be easily scaled without requiring expensive equipment and minimizing unnecessary congestion on the roads.**

I. INTRODUCTION

Researchers across the globe are trying to optimize traffic light systems to reduce carbon emissions and economic losses. There are two types of control used for traffic light system. One is conventional, where a preset time is set according to the average traffic on the road per day. As the green time schedule is fixed irrespective of the traffic, this often leads to congestion in lanes with the higher traffic flow. In this paper, the author proposes a GPS based adaptive method which dynamically calculates green light timings based on the traffic flow and prioritizes the lanes with the most traffic. When a user approaches an intersection, the mobile application sends the current and destination direction using mobile networks to an online central server and suggest appropriate lane the user. The online server then sends the data to the local traffic data to the micro-controller with GSM capabilities at the intersection. The micro-controller then calculates and updates the green light time of the intersection according to the traffic data and prioritizes the lane with higher traffic.

[1]Krishna Kant Sharma is a student of Department of EE, Delhi Technological University, India.

[2]Indu S is with the Department of ECE, Delhi Technological University, India.

II. RELATED WORK

Many researchers have been working on intelligent traffic light and traffic data collection systems. Authors of [1] proposed a data fusion approach to collect information about traffic in the critical urban road network. They collected the location, speed, and direction from GPS devices in vehicles being driven. Their aim was to use the data to improve traffic flow on the existing road network but they did not it use to manipulate the traffic light timings at road intersections. Authors of [2] proposed a video-based adaptive traffic signaling system to allocate fair green traffic times using data from a video camera and then assigning appropriate time using optimal weight calculator and fair weight calculator. The system is easy to implement but is expensive to scale as it requires installing a video camera at every intersection. This paper combines the work of the authors mentioned above to reduce traffic congestion and allocate fair green light time according to traffic flow. Author of [4] used detectors to the keep the count of arriving and leaving vehicles ans used it to extend or terminate the current signal phase. Author of [5] used on-board GPS unit on public transport vehicle to collect their speed and geographic location. The data is then sent to a backend online server which informs us about the traffic violations. Authors of [6] and [3] proposed a Big Data technology to promote the collection of users' data and utilizing it to reduce traffic congestion in urban road network. Author of [7] used embedded technology in vehicles to collect GPS data and pushing it to Traffic Monitoring System using GSM/GPRS. While authors of [8] used sensors and traffic servers to collect local traffic data and use it to control local traffic in the region. Although, the work proposed by authors of [7], [8] and [2] is effective, it is expensive to implement on a large scale. Big Data and data fusion are only the promising solutions which can be implemented on a large scale.

The previous works related to intelligent traffic scheduling the system also did not incorporate the route steps which the vehicle will take to reach its destination which is crucial information in calculating fair green light timings and prioritizing lane departures. This results in the allocation of unfair green light timings which further increases the congestion in lanes. The main advantage over the current radar and camera systems is that the current system's flexibility and scalability as there no requirement of expensive external hardware required to

978-1-5386-7024-8/19/$31.00 ©2019 IEEE

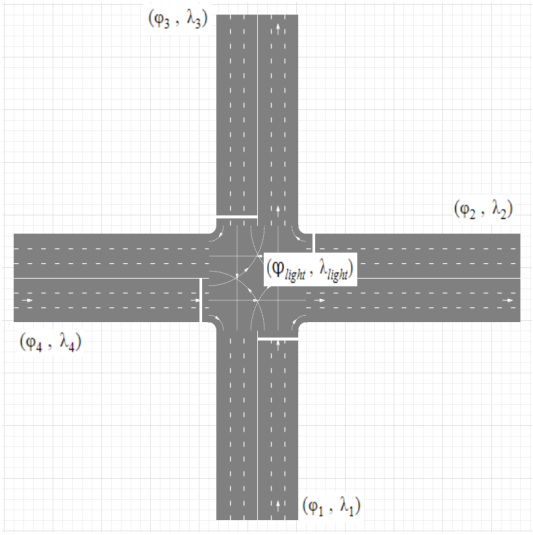


Fig. 1. Area of Intersection

install to collect traffic data.

III. PROBLEM DESCRIPTION

The road traffic lights' timings are preset according to average traffic on road per day. This does not incorporate the dynamic nature of the traffic. Due to the lack of dynamics in the existing traffic light system, delays and congestion are common problems because of the assignment of same green light timings irrespective of the traffic density of the lane. In cities like New Delhi, the situation of imbalanced traffic flow is rather common because working population resides in nearby cities like Noida and Gurugram. The imbalance in traffic is usually in the mornings and evenings when people travel to their work and residency.

IV. PROPOSED SOLUTION

We propose a GPS based adaptive traffic light timing and lane scheduling methodology which incorporates the varying nature of traffic. A mobile application sends the destination location, directional steps and current driving direction of the car driver which are in the near specific intersections using Google's Direction and Geocoding APIs. The data is collected in an online server, which then sends the traffic data to the GSM-enabled micro-controller at the intersection. The micro-controller calculates the green light time for each lane phase and allocates it to the traffic lights of the intersection.

V. METHODOLOGY

A. Marking Intersections' Geolocation

We mark the location of each intersection in the city with geolocation $(\phi_{light}, \lambda_{light})$, where ϕ_{light} is the latitude and λ_{light} is the longitude coordinate of the center of the intersection. Each intersection has been assigned a unique identity which helps to manage different intersections in the urban road network.

B. Defining Area Of Intersection (AOI)

It is the area which surrounds the intersection. The traffic in this area will only be considered in calculating the green light timings for each lane phase. AOI will be marked on all the four roads of intersection with geolocation (ϕ_i, λ_i), where ϕ_i is the latitude and λ_i is the longitude of the i^{th} corner and i = 1, 2, 3 and 4 with the intersection as the center. In this area, the mobile application notifies the driver to switch lanes according to his next direction step provided by the Google's Direction API, e.g., if the next direction of the driver is left, then the application will notify the driver to keep the leftmost lane when in AOI. This reduces the congestion and delays at the intersection further as one of the reasons for it is the wrong lane choices made by drivers. The intersection is set as the middle of the AOI square and with geolocation marked as $(\phi_{light}, \lambda_{light})$. Fig. 1 shows AOI of '+' intersection.

C. Defining Traffic Flow

Here, we show the more effective use of the road network by using lane pairs. We describe this methodology by using an example of 3 lane intersection as shown in Fig. 2. The driver has 3 options, i.e., straight, left turn and right turn. Fig. 3 shows 3 entries to the intersection. As this model is simulated for left-hand drive, the right-hand side is always free to turn and hence is not considered for a separate traffic light schedule. So, we are left only with 2 directions to consider in our system, i.e., straight and left. In a single '+' intersection, the traffic flow is structured in 6 phases. Fig. 4 shows all possible ways in which the traffic can flow in 6 phases. The first phase depends on the traffic of lane pairs, i.e., the lanes will be open for either north cars taking the turn and going straight or east cars taking the turn and north cars going straight. The one with more traffic is chosen as the first phase. If the traffic comes out to be same for both, the default phase is chosen as shown in Fig. 5. Similarly, this analogy can be extended to third, fourth and fifth phase. It can be represented in equation form as (2), (4), (5) and (7).

The second phase also depends on the traffic of lane pairs, i.e., the lanes will be open for either north and south cars taking turn or north and south traffic going straight. The one with higher traffic is chosen as the respective phase. If the traffic comes out to be the same for both, the default phase is chosen as shown in Fig.5. Similarly, this analogy is extended for the fifth phase. Both the second and fifth phase is represented in (3) and (6) respectively.

D. Collection of Traffic Data

As we need to collect the traffic data of the intersection, we develop mobile application in which the users can feed in their destination location and it will guide them to their destined location. As we use Google's Direction and Geocoding API, we have access to the user's current location and their respective direction steps for the commute. Geocoding API will provide us the current direction and whether the user is near the intersection or not.

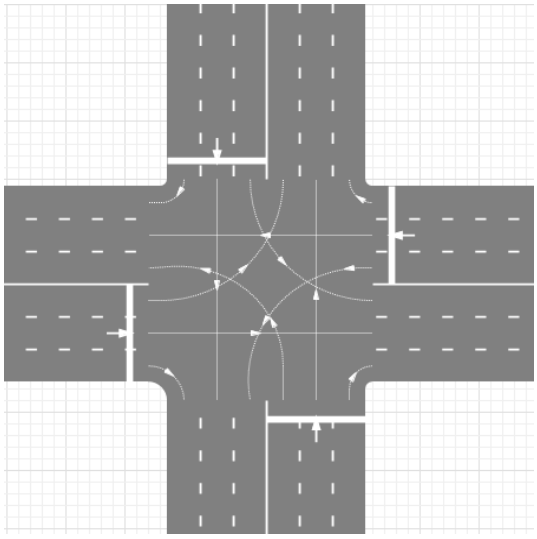


Fig. 2. 3 Lane '+' Intersection.

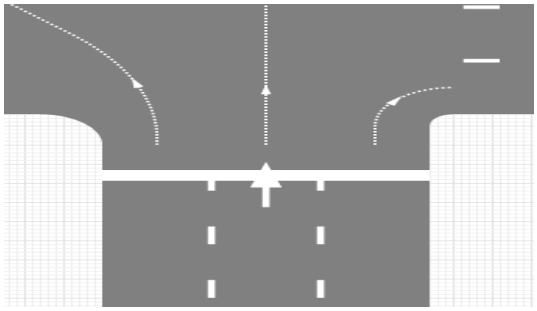


Fig. 3. Three entries to the intersection for one road.

The Direction API will provide us the directions that the user will take throughout the commute. The Direction API also tells us that whether the user has taken a successful direction step at the intersection or not. We store the database of the intersections of urban road networks and their respective AOI locally in the mobile application so that we can easily track the user and the intersection.

E. Online Central Server for Data Collection

The Online Central Server keeps track of all the cars which are in the AOI of the respective intersections. When the driver enters the AOI of an intersection, the mobile application pushes the user data to the online central server. The server keeps track of the number of cars in a particular direction and their respective next direction for the specified intersection using (1) and then pushes the data to the GSM-enabled micro-controller for the respective traffic light. Whenever the traffic in a particular phase becomes zero, the online server pushes the command to skip the current phase.

$$X_{st} := X_{st} + 1. \tag{1.a}$$

, for cars which will go straight.

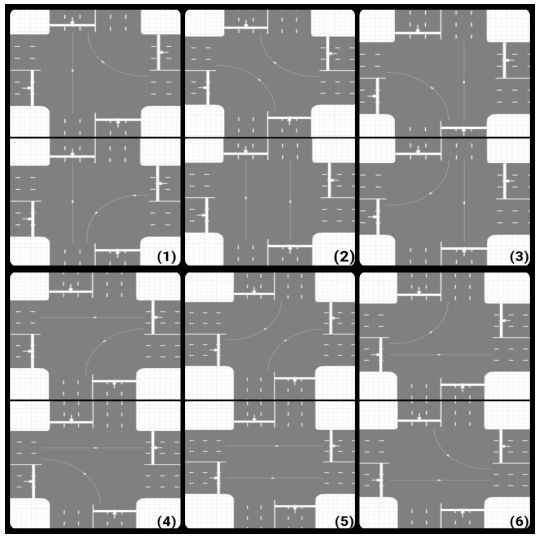


Fig. 4. All possible lane phases

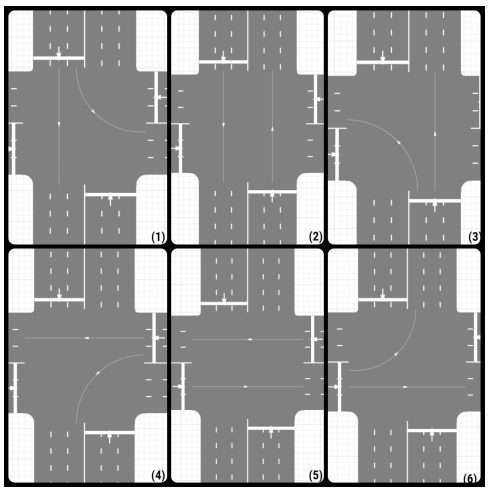


Fig. 5. Traffic Phases In Default Mode

$$X_{turn} := X_{turn} + 1. \tag{1.b}$$

, for cars which will take a turn.

$$X_{st} := X_{st} - 1. \tag{1.c}$$

, for cars which have successfully gone straight.

$$X_{turn} := X_{turn} - 1. \tag{1.d}$$

, for cars which have successfully taken a turn.

F. Hardware Architecture and implementation in Traffic Lights

The traffic data is received from the Online Central Server using GSM-enabled Arduino module. The Arduino microcontroller then computes the green light time for each lane pair and updates it after every lane pair phase so as to make the system updated with the current traffic situation. Fig.6 shows the whole process of traffic data collection and distribution among different intersections.

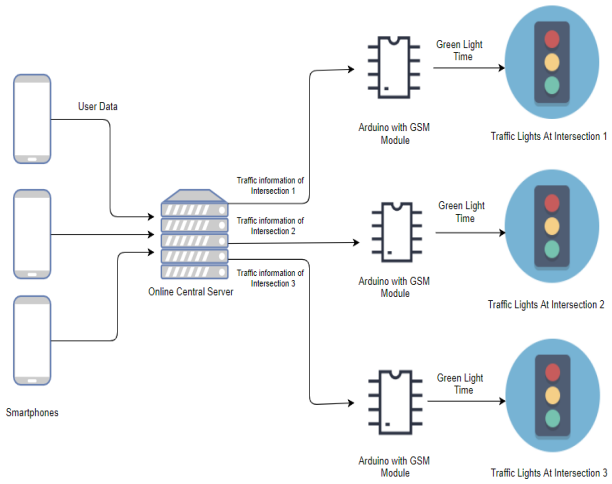


Fig. 6. Collection and distribution of traffic data.

G. Green Light Time Calculation

We calculate the number of cars in each lane pair phase using (2), (3), (4), (5), (6) and (7) and allocate lane pairs according to current traffic.

After calculating the N_i for each phase, we allocate green time to each phase using (10), where N_i is the number of cars in i^{th} phase. Hence, the lane with the most traffic will be allotted the more green light time. The time is updated after each lane pair phase is completed. The traffic light also skips phase whenever the traffic in the phase is cleared before its green time is over or whenever the phase is clear.

When the traffic is clear in all lanes, the traffic lights will be assigned equal green light timings and the phases will run in the default phase order as shown in Fig. 4.

$$N_1 = \begin{cases} N_{st} + N_{turn}, & \text{if } N_{st} + N_{turn} \geq N_{st} + E_{turn} \\ N_{st} + E_{turn}, & \text{otherwise.} \end{cases} \tag{2}$$

$$N_2 = \begin{cases} N_{st} + S_{st}, & \text{if } N_{st} + S_{st} \geq N_{turn} + S_{turn} \\ N_{turn} + S_{turn}, & \text{otherwise.} \end{cases} \tag{3}$$

$$N_3 = \begin{cases} S_{st} + S_{turn}, & \text{if } S_{st} + S_{turn} \geq S_{st} + W_{turn} \\ S_{st} + W_{turn}, & \text{otherwise.} \end{cases} \tag{4}$$

$$N_4 = \begin{cases} E_{st} + E_{turn}, & \text{if } E_{st} + E_{turn} \geq E_{st} + S_{turn} \\ E_{st} + S_{turn}, & \text{otherwise.} \end{cases} \tag{5}$$

$$N_5 = \begin{cases} E_{st} + W_{st}, & \text{if } E_{st} + W_{st} \geq E_{turn} + W_{turn} \\ E_{turn} + W_{turn}, & \text{otherwise.} \end{cases} \tag{6}$$

$$N_6 = \begin{cases} W_{st} + W_{turn}, & \text{if } W_{st} + W_{turn} \geq W_{st} + N_{turn} \\ W_{st} + N_{turn}, & \text{otherwise.} \end{cases} \tag{7}$$

We normalize N_i by using (8).

$$N_{i(norm)} = \frac{N_i}{S}, \text{ i = 1,2,...,6} \tag{8}$$

Where,

$$S = \sum_{i=1}^{6} N_i \tag{9}$$

After calculating the N_i for each phase, we allocate green light timings to each phase using (10). Hence, the lane with the most traffic will be allotted the more green light time. The time is updated after each lane pair phase is completed. The traffic light skips phase whenever the traffic in the phase is cleared before its green time is over or whenever the phase is clear. If the calculated t_i is less than a minimum value set according to the particular intersection, then t_i is set to zero and the respective phase is skipped.

$$t_i = N_i * T, \tag{10}$$

where t_i is the allocated time to i^{th} phase. When the traffic in all lanes is clear, the green light time will be assigned equal timings and the phases will run in the default phase order as shown in Fig. 4.

VI. Simulation

According to the proposed design, the simulation of a single '+' intersection was made. The simulation was made using AnyLogic [9] to simulate realistic driver's behavior and car characteristics. Thus different traffic scenario were created and finally the green light times assigned to each lane pair and the average traffic over the course of an hour were observed. A custom car agent was created in the simulation with 'currentDir' and 'nextDir' as parameters assigned to each car which controls the current and the next direction of the car respectively. When the car reaches AOI, it changes lane according to the next direction assigned to it. In the simulation of the model, we can control the following parameters:

- The number of car flowing per hour on the road.
- The rate of the car with specific next direction.
- Total time cycle of Intersection.

In the first simulation, we take the same intersection as shown in Fig. 1. We set the total time of the cycle (T) to 480 sec. To simulate real-life scenario, where the traffic is distributed unequally, we set rate (cars/hour) of the total number of cars coming from the West to 900, coming from the South to 450, coming from East to 500 and coming from North to 600. Table I shows the distribution of the rate of cars with specific next turns. Fig. 7 shows a scenario while the simulation is running. The simulation is run for an hour to get a better analysis of the proposed solution. Table II compares the green light time of conventional green light time with our calculated green light time for each phase for the given scenario. Fig. 8 shows the number of cars on each road during one hour of simulation in AOI when conventional timings are used. Fig. 9 shows the number of cars on each road during one

hour of simulation in AOI when GPS based traffic light scheduling is used. The results shown in Table II. show that our proposed solution prioritizes the lane pairs with the most traffic. On comparing Fig. 8 and Fig. 9, we can see that our proposed solution reduces the congestion in roads with most traffic by evenly distributing it among the other roads. The average number of cars present in AOI of both Fig. 7 and Fig. 6 are compared in Table III.

TABLE I

DISTRIBUTION OF THE RATE OF CARS WITH SPECIFIC NEXT DIRECTION.

Road	Total (cars/hour)	Straight (cars/hour)	Left (cars/hour)	Right (cars/hour)
North	600	200	200	200
South	450	150	150	150
East	500	300	100	100
West	900	450	225	225

TABLE II

AVERAGE NUMBER OF VEHICLES ON ROAD DURING SIMULATION.

Road	Average number of cars (Conventional)	Average number of cars (GPS based Adaptive scheduling)
North	25.45	21.93
South	11.99	19.25
East	19.07	19.31
West	60.75	47.93

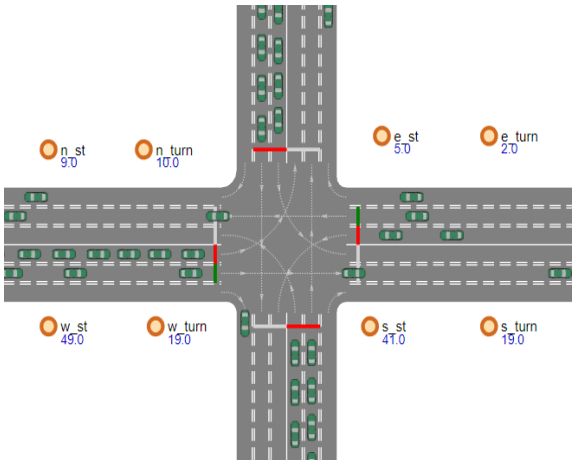


Fig. 7. Traffic flow in AOI in the fifth phase.

In the second simulation, we take the same intersection as shown in Fig. 1. We set the total time of the cycle to 480 sec. We set rate (cars/hour) of the total number of cars coming from the West to 400, coming from the South to 900, coming from East to 500 and coming from North to 600. Table IV shows the distribution of the rate of cars with specific turns. Here, the majority of cars are going to the north direction. Fig. 10 shows a scenario while the simulation is running. Table V compares the green light time of conventional green light time with our calculated green light time for each phase. Fig. 11 shows the number of cars on each road during one hour of simulation in AOI when conventional timings are used. Fig. 12 shows the number of cars on each road during one hour of simulation in AOI when calculated green light timings are allocated. The results shown in Table V show that our proposed solution prioritizes the lane pairs with the most traffic. On comparing Fig. 9 and Fig. 10, we can see that our proposed solution reduces the congestion in roads with most traffic by evenly distributing it among other roads. The average number of cars present in AOI of both Fig. 10. and Fig. 9. are compared in Table VI.

TABLE III

DISTRIBUTION OF THE RATE OF CARS WITH SPECIFIC NEXT DIRECTION.

Road	Total (cars/hour)	Straight (cars/hour)	Left (cars/hour)	Right (cars/hour)
North	600	200	200	200
South	900	600	150	150
East	500	300	100	100
West	400	100	200	100

TABLE IV

AVERAGE NUMBER OF VEHICLES ON ROAD DURING SIMULATION.

Road	Average number of cars (Conventional)	Average number of cars (GPS based Adaptive scheduling)
North	9.39	10.58
South	103.64	49.58
East	19.21	23.07
West	21.57	17.35

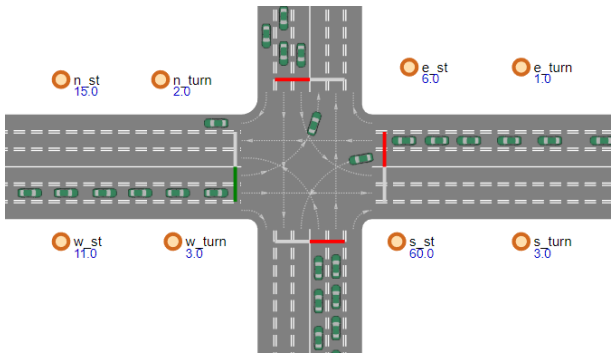


Fig. 8. Traffic flow in AOI in the sixth phase.

TABLE V

GREEN LIGHT TIME FOR THE INTERSECTION IN FIG. 7.

Phase Number	Lanes open in each phase (Conventional)	Green Time (in seconds) Conventional	Lanes open in each phase	Green Time (in seconds)	Number of vehicles in prioritized lanes
1	N to S and N to E	60	N to S and N to E	33.16	19
2	N to S and S to N	120	N to S and S to N	87.27	50
3	S to N and S to W	60	S to N and S to W	104.72	60
4	E to W and E to S	60	E to W and S to W	41.89	24
5	E to W and W to E	120	E to W and W to E	94.25	54
6	W to E and W to N	60	W to E and W to N	118.69	68

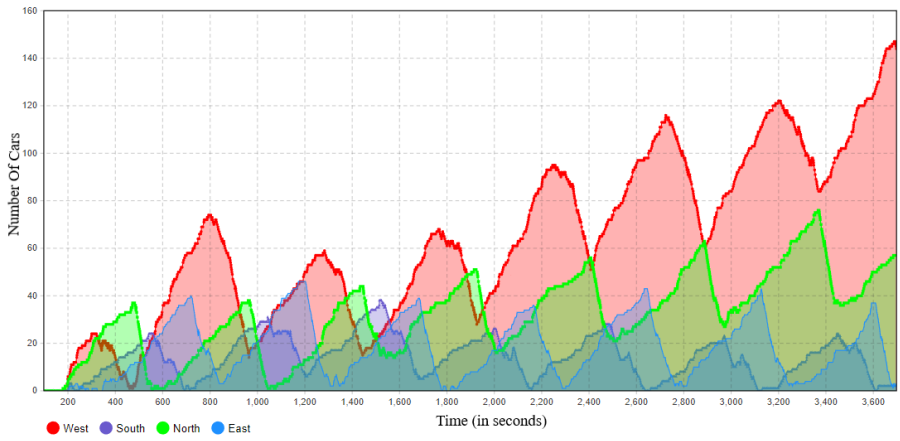


Fig. 9. Time versus Traffic on each road in AOI with conventional green light scheduling.

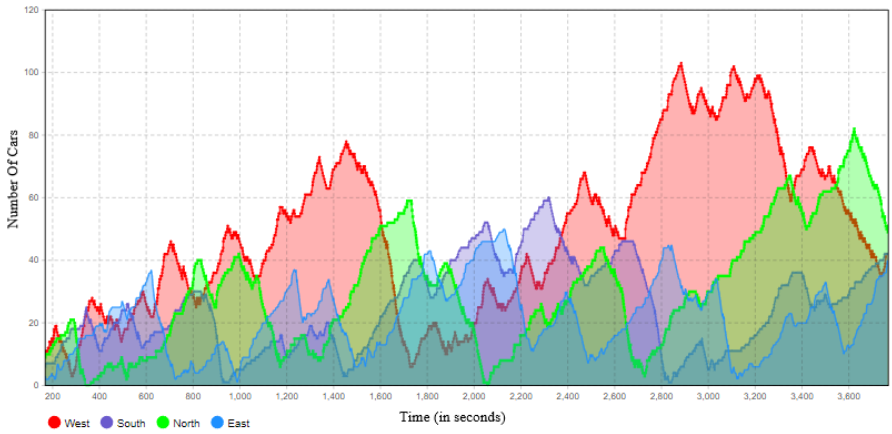


Fig. 10. Time versus Traffic on each road in AOI with GPS based adaptive scheduling.

TABLE VI

GREEN LIGHT TIME FOR THE INTERSECTION IN FIG. 7.

Phase Number	Lanes open in each phase (Conventional)	Green Time (in seconds) Conventional	Lanes open in each phase	Green Time (in seconds)	Number of vehicles in prioritized lanes
1	N to S and N to E	60	N to S and N to E	41.85	17
2	N to S and S to N	120	N to S and S to N	184.61	75
3	S to N and S to W	60	S to N and S to W	155.07	63
4	E to W and E to S	60	E to W and S to W	22.15	9
5	E to W and W to E	120	E to W and W to E	41.85	17
6	W to E and W to N	60	W to E and W to N	34.46	14

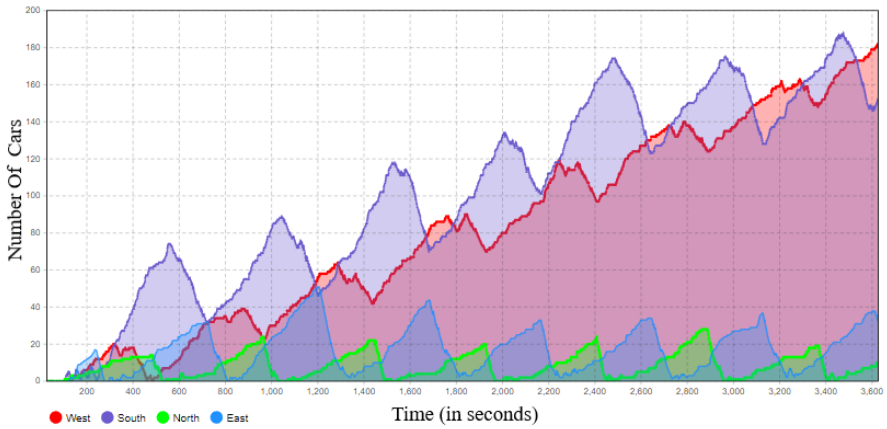


Fig. 11. Time versus Traffic on each road in AOI with conventional green light scheduling.

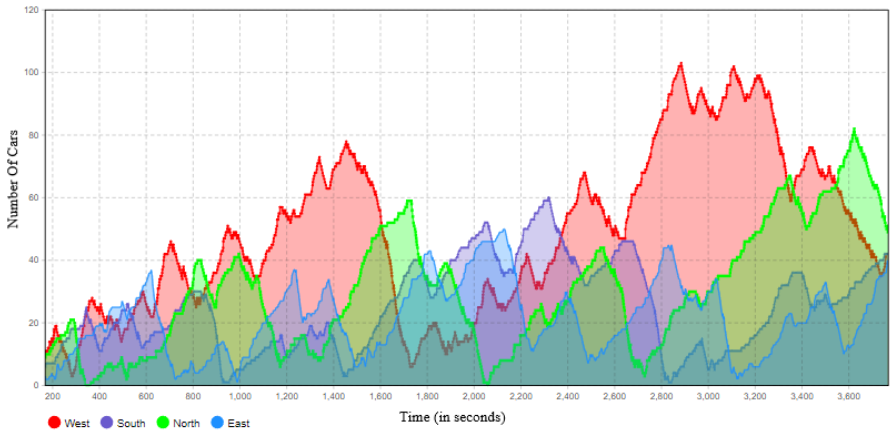


Fig. 12. Time versus Traffic on each road in AOI with GPS based adaptive scheduling.

VII. Conclusion and Future Work

This paper has presented a GPS based dynamic traffic light timings and lane scheduling which reduces the congestion of the roads with higher traffic flow by prioritizing the respective road with the most traffic and allocating a longer green light time. The author compares the proposed solution with the conventional traffic light scheduling system using AnyLogic simulator. The results indicate that the proposed solution allocates dynamic timing schedule and lane pairs for different traffic conditions and significantly reduces the congestion in the roads with higher traffic flow.

However, applying this methodology in real life is still a challenging task. In our future work, we will working on the developing the hardware and mobile software and testing our system in a small road intersection.

References

[1] Tiedong W. and Jingjing H. (2014), "Applying floating car data in traffic monitoring", *2014 IEEE International Conference on Control Science and Systems Engineering.*

[2] Indu S., Nair V, Jain S., Chaudhury S. (2013), "Video based adaptive road traffic signaling", *2013 Seventh International Conference on Distributed Smart Cameras (ICDSC).*

[3] Tanwar R., Majumdar R., Sidhu, G. S. and Srivastava, A. (2016), "Removing traffic congestion at traffic lights using GPS technology", *2016 6th International Conference - Cloud System and Big Data Engineering (Confluence).*

[4] Lin Y. (2014), "The Design and Simulation of Intelligent Traffic Signal Control System Based on Fuzzy Logic", *Foundations'of Intelligent Systems, 965âĂŞ973.*

[5] Tarapiah S., Atalla. S., Muala, N. and Tarabeh S. (2014), "Offline Public Transportation Management System Based on GPS/WiFi and Open Street Maps", *2014 Sixth International Conference on Computational Intelligence, Communication Systems and Networks.*

[6] Liu Y (2018),"Big Data Technology and Its Analysis of Application in Urban Intelligent Transportation System", *2018 International Conference on Intelligent Transportation, Big Data and Smart City (ICITBS).*

[7] Liping. X., and Dangying L. (2014), "A Novel Embedded Vehicle Terminal for Intelligent Transportation", *2014 Fifth International Conference on Intelligent Systems Design and Engineering Applications.*

[8] V. K. Verma and R Choudhari and S. K. Singh and T. M. Abhai Pratap Singh and P. Srivastava, "Intelligent transport management system using sensor networks", *IIVS, pages 4-6, June 2008.*

[9] https://www.anylogic.com/

GST IN INDIAN ECONOMY AND ITS POSSIBLE ROBUSTNESS

Mehul Gupta, Delhi Technological University
Prashant Sharma, Delhi Technological University

ABSTRACT

The Goods and Service Tax bill applied uniformly across the nation, a form of a levy on the manufacture, sales, and consumption of goods and services. This paper attempts to highlight the performance of the implementation of GST against the previous tax structure and a hypothetical case for checking the efficiency of GST during unexpected scenarios to the economy by utilizing GTMA using the scale of satisfaction index and permanent functions to compute a resultant overall satisfaction index. This paper discusses the GST on the scale for the ease of the market to incorporate GST and its performance against the specific factors.

Keywords: GST, GTMA, Macroeconomics, Taxation, Government Revenue, Macroeconomic Stabilization

INTRODUCTION

What is GST?

GST, also known as Goods and Service Tax, passed in the Lok Sabha on 29th March 2017 and came into effect from 1st July 2017 under the GST Act or Constitution (One Hundred and First Amendment) Act, 2016. The tax went into the picture as another type of indirect tax, which is a levy in place of previous existing taxes. The idea behind the introduction of is to have GST replaced an array of indirect taxes under a unified tax regime and is thus is expected to reshape the country's economy anew.

History of GST

The L K Jha committee was foremost dyed the need to turn into the Value Added Tax regime (VAT) in India, 1974. Chelliah committee proposed VAT or GST execution in 1991. Service Tax executed in India, 1994. In 2000, the Vajpayee Government started the discussions on GST, by setting up an empowered committee, headed by Asim Dasgupta, the then Finance Minister, Government of West Bengal. The committee assigned the task of designing the GST model and overseeing the IT backend preparedness for its rollout. Moving towards GST was then again reflected by, then, Union Finance Minister in his Budget discourse for 2006-07. At first, it was recommended that GST would be presented from first April 2010. The Empowered Committee of State Finance Ministers (EC), which had figured the outline of State VAT was asked to create a guide and structure for GST for altering according to the need of everyone to be catered. Joint Working Groups of authorities having agents of the States and additionally the center were set up to inspect different parts of GST and draw up reports particularly on exclusions and edges, tax collection of administrations and tax assessment of between State supplies in accordance to fulfill the role in removing excess taxes on multi-level stages. In light

of dialogs inside and amongst it and the Central Government, the EC discharged its First Discussion Paper (FDP) on the GST in November 2009. The FDP released in November 2009 beckons out highlights of the proposed GST and had framed the reason for exchange between the Centre and the States up until recent. The Goods and Services Tax (GST) is a novel idea that simplifies the complex tax structure by supporting and enhancing the economic growth of a country. GST is an extensive tax levy on the manufacturing, sale, and consumption of goods and services at a national level Dani (2016).

The Goods and Services Tax Bill or GST Bill, also referred to as The Constitution (One Hundred and Twenty-Second Amendment) Bill, 2014, initiates a Value-added Tax to be implemented on a national level in India. GST will be an indirect tax at all the stages of production to bring about uniformity in the system. On bringing GST into practice, there would be an amalgamation of Central and State taxes into a single tax payment. It would also enhance the position of India in both domestic as well as international market. At the consumer level, GST would reduce the overall tax burden, which is the current estimate at 25-30% Dani (2016). Under this system, the consumer pays the final tax, but an efficient input tax credit system ensures that there is no cascading of taxes- a tax on tax paid on inputs that go into the manufacture of goods. To avoid the payment of multiple taxes such as excise duty and service tax at the Central level and VAT at the State level, GST would unify these taxes and create a uniform market throughout the country. The integration of various taxes into a GST system will bring about an effective cross-utilization of credits. The current system taxes production, whereas the GST will aim to tax consumption. Experts have enlisted the benefits of GST as under:

1. It would introduce a two-tiered One-Country-One-Tax regime.
2. It would subsume all indirect taxes at the center and the state level.
3. It would not only widen the tax regime by covering goods and services but also make it transparent.
4. It would free the manufacturing sector from the cascading effect of taxes, thus improving the cost-competitiveness of goods and services.
5. It would bring down the prices of goods and services and, thus, increase consumption.
6. It would create a business-friendly environment, thus by increase the tax-GDP ratio.
7. It would enhance the ease of doing business in India.
8. Some of the best GST systems across the world like Singapore and New Zealand use single GST, but India has opted for a dual GST model Dani (2016) &Venkatesh (2015) The change in tax structure is expected to have a significant impact on the supply chain in India. A comparison of the GST rates of some countries with the proposed India GST rate is mentioned in Table 1 Venkatesh (2015).

TABLE 1
RATES OF TAXES LEVIED IN DIFFERENT COUNTRIES

Country	The rate of GST (%)
Australia	10
Canada	5
New Zealand	15
Singapore	7
Malaysia	6
Sweden	25
India	27*

(*Maximum levy on goods applicable)

LITERATURE REVIEW

GST as a tax structure can be loosely defined as the amalgamation of several centers and state-driven taxes into a single umbrella of the GST tax regime. The primary driving factor for an introduction to GST has created an even market for making tax structure easier at domestic level and international investment favourable environment for tax-driven Indian Economy. At the consumer level, GST would reduce the overall tax burden, which is currently estimated at 25-30% Dani (2016). In the face to create tax structure unified and straightforward to reduce the problems of payments of a different centre run and state-run taxes at different levels of the supply chain, it dissolves in all the taxes into a single one to create an unformalized market with easier access. While it has been noted that most of the tax will be burdened over the end-user and consumers, but GST still holds the claim to reduce total tax burdens by margins. However, the devised bill may have been drafted with the idea to ease in the current economy and create a tax-friendly market for the producers. The idea of ease of understanding and to uniform the tax structure can be structured on various factors and parameters like ease of application of tax, both at corporate and individual level, and reducing the confusion at various levels of taxation, which may also lead to tax evasions as well. There still can be several factors on which it can be measured by its performances over the term of implementations, which can be significantly varied with high correlation to each other and while some may be entirely independent of each other. This type of approach to determine the performance a particular strategy on selected varied factors requires checking the performance of the observed scenarios by each factor independently and simultaneously at a level dependent on one another. This type of evaluation can give hidden information about the structure which can be much clouded on the terms of factors when seen as a whole. The method described also gives the picture of the nature of planning during the stages of the drafting of such strategic steps and the agenda of the implementation body about the type of problems it was tailored for, both at bureaucratic and corporate levels. The bill can be further tested by satisfaction of individuals towards its implementation Moorman et al. (1993). Here the different dispositions of different factors and attributes are being observed by the satisfaction index to find the overall performance of the cases in different aspects of the uniformity and prospects of future elevation. The three scenarios that have been studied in this case are current implementation and its implementation period, Post-Demonetisation period and one of the hypothetical cases if the GST would have been introduced following demonetization simultaneously. The prior two cases are real-world instances and in their accordance, while the latter one is hypothetical, which aims to find the robustness of GST as a bill for Indian domestic and international markets in cases of sharp blows. The three given disposition alternatives are chosen and are then tested for their implementations on the specific factors. The three cases are introduced in brief below.

Current Implementation or Post-GST Period

The case illustrates the period after the successful implementation of GST in India. The basis of comparison is the Pre-GST period with the conventional tax-regime that was present in

India. The Lok Sabha passed the Bills on 29th March 2017. The Rajya Sabha passed these Bills on 6th April 2017 and was then enacted as Acts on 12th April 2017. For the sake of convenience, the case is referred to as Post-GST period later in the paper.

Post-Demonetisation Period

This case illustrates the condition of domestic Indian markets and trade in the period after the demonetization was implemented on the existing currency notes by then Prime Minister Narendra Singh Modi. The announcement of demonetization was made on 8th November 2016 with the effect of denouncing all high currency notes of ₹1000 and ₹500 and getting new currency notes of ₹2000 and ₹500 in circulation in hope to curb the black money and counterfeit currency.

Post-Demonetisation period with successively implemented GST (Hypothetical)

This case is taken hypothetically for the instance if the GST has happened after just weeks to demonetization. The primary aim of this case is to check the robustness of the drafted bill and for the understanding its validation in case of any such unexpected future occurrences to float out the country's economy with the least citizen discomfort.

The process of selection for the cases mentioned above and alternatives can vary according to the observer, but the main reason for choosing the above-specified alternatives for the selection can be mainly attributed to the fact them being the most highlighted events in the Indian economy in recent periods. The focus of the project is on the application of GST and its performance measure for the soundest effect of it on the then Indian markets and with more considerable implications on the current and future trade of the domestic and foreign market affairs. The second case of the Demonetisation was taken into consideration with the fact that India observed a significant slowdown after the implementation in the financial year 2017. This case works as a precise scale for any of the unexpected slowdowns that may occur in the future to provide a reference. Some of the ramifications of the same can be said to be rippled during the global slowdown in late 2019 global economic slowdown as India continues to face declining GDP growth. The factors on which the above cases are grounded upon and are compared their performance against are mainly inclined towards the ease of working in Indian markets while others are in conjugation to the performance of the economy. While the significant setbacks were visible on the comfort of the citizens as cited by all major news outlets, the effects of the demonetization on the market and trade affairs remain unclear, with many of the economists and finance experts divided on their views. The last case is a hypothetical one where the GST introduced in the period of the near demonetization implementations. While the GST bill was introduced in early July 2017 to be efficiently implemented in the markets, the case of demonetization has already been in vain with its significant impacts already delivered to the Indian markets. Thus it has been assumed that if the GST was successfully implemented in the markets in the period of late November or early December 2016. The primary purpose of the case is to find out the performance of GST in the cases of uncertainties and precipitous effects of specific instances.

The GST bill is being scaled for the robustness on the basis that whether it can help as a deterrent for the future instances where any uncertain event with large-scale implications can be withstood or if the structure of Indian economy and taxes become more brittle and vulnerable. The main objective of the study is found to observe the GST as a bill for the stability of Indian domestic and foreign markets. Pierre J. Richard et al. (2009) developed methods to calculate the organizational performance using multi-dimensional conceptualization Agrawal et al. (2016). Used the methods of GTMA for the analysis of different alternatives and their feasibility on the scale of specific features for the topic of reverse logistics.

METHODOLOGY

The comparison is performed on different levels of influence subjected to the discussed factors by their inclusivity on the drafting of the act. The draft in itself is devised on some factors which are to be concerned for significant welfare on the scale for the majority of the populace and future aspect of amendment and execution ability. The approach towards the comparison of these different factors to find the inclination of the executed GST and future aspects for the improvement to attain stability and prospects for initiating a smoother globalization trend for the Indian economy. Several of the factors are to be considered for the selection of an approach to achieve nearer to actual results based on the collected facts and evidence. In case the factors thought on which the study is scaled are in total disjoints Rao & Padmanabhan (2007) to each other, approaches such as TOPSIS and AHP are selected algorithms. However, the factors suggested are not in total disjoints to each other about a certain level of influence upon each other. Thus the approach proposed is not suitable for the same. DEA Toloo (2011) algorithm is suitable for higher accuracy but requires large computations and deviates profoundly from the accurate answer for more numbers of factors. ANP Yang (2011) fails in cases of hierarchical relationships for the computations. Thus the method used for the computations selected is Graph Theory and Matrix Approach (GTMA) Agrawal et al. (2016), which is not bound by any limitations mentioned previously above.

The GTMA approach allows for analysis of highly directed graphs with many nodes in place, and the complexity of the graph is high, to draw any useful results from the actual visualizations. Many software allows for the matrix approach application for handy and easy computations. This particular approach is chosen for its compatibility with any number of nodes and attributes. The concept fully characterizes the considered selection problem, as it contains all possible structural components of the attributes and their relative importance Rao (2007). In the case of even small deviations in permanent functions may lead to significant changes in satisfaction index, and thus allowing for natural ranking of factors for in the descending order of satisfaction or ascending order of index of places with necessary improvement.

Further, these rankings can be used to find the relative influence each of them may hold on others to devise natural executing amendments for maximizing results. Because of all these advantages discussed above, the proposed study has applied GTMA for the selection of disposition alternatives. The examples of applications of GTMA in fields of vast technology and

finance results to draw various conclusions. Jain et al. (2016) used the process to create a model to analyze the performance variables of a flexible manufacturing system (FMS). RV Rao (2007) used the principal of GTMA for analysing Operational Performance Evaluation of Competing Companies. Wagner et al. (2010) used the same for assessing the vulnerability of supply chains using graph theory Muduli et al. (2013) used the algorithm to model and analysed identified and ranked the barriers to the green supply chain in the Indian mining industry.

The main elements of the GTMA model are the elements of the graph, i.e., Nodes and Edges. Each edge act to form a directional vector connecting two nodes depicting the nature of interactions between the two nodes in the observation at the time. Each node is a representation of a single attribute or factor in the discussion. The directional graph (or digraph) is prepared, which is, in turn, used to understand the relations between each attribute or factor which is capable of influencing the satisfaction index. Numbers of nodes (M) in the digraph are taken to be equal to the number of attributes considered for the study, and arrow or directed edge connecting two nodes represents their relative importance. A systematic approach, based on the method discussed by Malik (2015), is developed for the study.

Any digraph consists of particular number of nodes N = $\{n_i\}$, where *i* is set of numbers {1 , 2 , 3 , ….. , M} and the set of directed Edges E = $\{x_{ij}\}$) Agrawal et al. (2016) . A node n_i represents an i^{th} attribute for the selection of satisfaction alternative, and edges represent the relative importance of these attributes. For example*, if the attribute "i" is more important than the attribute "j" while comparing two attributes for their relative preference, then the vector or arrow or edge is directed from node "i" to node "j." The value of preference is represented by* x_{ij}. If the attribute "*j*" is more important than the attribute "*i*," then arrow or edge is directed from node "*j*" to node "*i*." The value of preference is represented by x_{ji}. The function of the digraph is to create a visualization of the attributes and their relative preference to each other to help them analyze quickly. However, in case of the number of nodes or attributes increases then the visualization of such a graph may not deliver relevant results due to highly complex structures, thus in such cases where the number of nodes exceeds from the certain value after which the digraph may become complex the whole solution is converted into a matrix.

The matrix formed is a type of transformed digraph in the representation only. The matrix formed is **Relative Importance Matrix** Slack et al. (1994) where all off-diagonal elements discuss the relative preference of one attribute to the other. The diagonal elements give relative satisfaction or improvement alternative. Since the digraph or matrix is variable changing with the number of the attribute in observation, a standard matrix function is devised for the same. The function is called *Permanent Function* is computed rather than the determinant of the matrix. The mentioned function taken is because the determinant function of the matrix if computed may accompany the loss of information due to carrying negative signs. However, the complete information is considered in Permanent Function thus is wholly used as preferred Rao et al. (2007).

The study selected attributes are:

1. Ease of Filing (EoF).
2. Government Incentives (GI).
3. Foreign Trade Barriers (FTB).
4. Total Monetary Payment (TMP).
5. Customer Friendly (CF).
6. Availability of Loans (AoL).
7. Procurement of Materials (PoM),
8. Supplier Management (SM),
9. Warehousing (Wh),
10. Credit Cash Flow (CCF).

The digraph for the discussed ten attributes is shown in Figure 1 Agrawal et al. (2016).

Since the representation of the digraph is complicated to decode from the visual interpretation of Fig 1, thus a matrix [X] has been developed for the more fundamental analysis of the same. The matrix produced is an M × M type of matrix for ten different attributes consisting of the diagonal elements. The diagonal elements are represented by $Z_{ij,}$ where i=j and in case of off-diagonal elements are represented by x_{ij} where $i \neq j$.

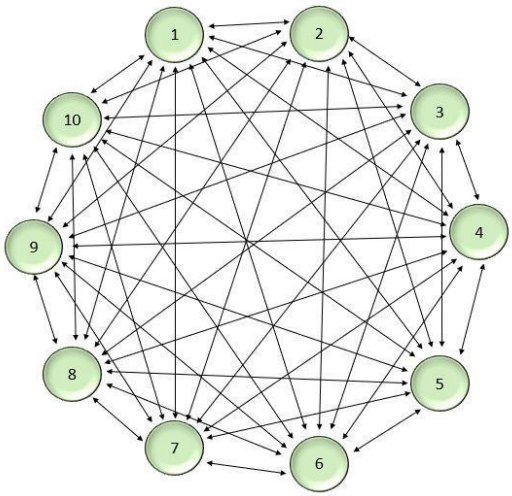


FIGURE 1
EXPECTED DIGRAPH OF SUGGESTED INTER-RELATIONSHIP AMONG DIFFERENT ATTRIBUTES

The off-diagonal elements describe the relative satisfaction of each item and are the same for all alternatives or improvement alternatives. Diagonal elements indicate the need for improvement of an attribute for the given alternative and may differ for each choice.

$[X]$	EOF	GI	FTB	TMB	CF	AoL	PoM	SM	Wh	CCF
EOF	$Z1$	$X12$	$X13$	$X14$	$X15$	$X16$	$X17$	$X18$	$X19$	$X110$
GI	$X21$	$Z2$	$X23$	$X24$	$X25$	$X26$	$X27$	$X28$	$X29$	$X210$
FTB	$X31$	$X32$	$Z3$	$X34$	$X35$	$X36$	$X37$	$X38$	$X39$	$X310$
TMB	$X41$	$X42$	$X43$	$Z4$	$X45$	$x46$	$X47$	$X48$	$x49$	$X410$
CF	$X51$	$X52$	$X53$	$X54$	$Z5$	$X56$	$X57$	$X58$	$X59$	$X510$
AoL	$X61$	$X62$	$X63$	$X64$	$X65$	$Z6$	$X67$	$X68$	$X69$	$X610$
PoM	$X71$	$X72$	$X73$	$X74$	$X75$	$X76$	$Z7$	$X78$	$X79$	$X710$
SM	$X81$	$X82$	$X83$	$X84$	$X85$	$X86$	$X87$	$Z8$	$X89$	$X810$
Wh	$X91$	$X92$	$X93$	$X94$	$X95$	$X96$	$X97$	$X98$	$Z9$	$X910$
CCF	$X101$	$X102$	$X103$	$X104$	$X105$	$X106$	$X107$	$X108$	$X109$	$Z10$

The permanent function of the above matrix is:

$$\boldsymbol{Per}(\boldsymbol{X}) = \prod_{i=1}^{A} Zi + \sum_{i=1}^{A}\sum_{j=i+1}^{A} \dots\dots\dots\dots\sum_{A=s+1}^{A}(Xij.Xji).Zk.Z1.Zm.Zn \dots\dots Zs.ZA +$$
$$\sum_{i}^{A-2}\sum_{j=i+1}^{A-1}\sum_{k=j+1}^{A} \dots\dots\dots\sum_{A=s+1}^{A}(Xij.Xjk.Xki + Xik.Xji.Xkj).Z1.Zm.Zn \dots\dots Zs.ZA +$$
$$\sum_{i=1}^{A-3}\sum_{j=i+1}^{A}\sum_{k=i+2}^{A-1}\sum_{l=i+2}^{A} \dots\dots\sum_{A=s+1}^{A}(Xij.Xji)(Xkl.Xlk)Zm.Zn \dots\dots Zs.ZA +$$
$$\sum_{i=1}^{A-3}\sum_{j=i+1}^{A+1}\sum_{k=i+1}^{A}\sum_{l=i+1}^{A} \dots\dots\sum_{A=s+1}^{A}(Xij.Xjk.Xkl.Xli +$$
$$Xil.Xji.Xkj.Xlk)Zm.Zn \dots\dots Zs.ZA +$$
$$\sum_{i=1}^{A-2}\sum_{j=i+1}^{A-1}\sum_{k=j+1}^{A}\sum_{l=1}^{A-1}\sum_{m=l+1}^{A} \dots\dots\sum_{A=s+1}^{A}(Xij.Xjk.Xki +$$
$$Xik.Xji.Xkj)(Xlm.Xml)Zn \dots\dots Zs.ZA +$$
$$\sum_{i=1}^{A-4}\sum_{j=i+1}^{A-1}\sum_{k=i+1}^{A}\sum_{l=i+1}^{A}\sum_{m=j+1}^{A} \dots\dots\sum_{A=s+1}^{A}(Xij.Xjk.Xkl.Xlm.Xmi +$$
$$Xlk.Xji.Xkj.Xml.Xim)Zn \dots\dots Zs.ZA +$$
$$\sum_{i=1}^{A-5}\sum_{j=i+1}^{A-1}\sum_{k=j+1}^{A}\sum_{l=1}^{A-2}\sum_{m=l+1}^{A-1}\sum_{n=m+1}^{A} \dots\dots\sum_{A=s+1}^{A}(Xij.Xjk.Xki +$$
$$Xik.Xji.Xkj)(Xlm.Xmn.Xnl + Xln.Xml.Xnm)Zo \dots\dots Zs.ZA +$$
$$\sum_{i=1}^{A-5}\sum_{j=i+1}^{A}\sum_{k=i+1}^{A-3}\sum_{l=i+2}^{A}\sum_{m=k+1}^{A-1}\sum_{n=k+2}^{A} \dots\dots\sum_{A=s+1}^{A}(Xij.Xji)(Xkl.Xlk)(Xmn.Xnm)Zo \dots\dots Zs.ZA +$$
$$\sum_{i=1}^{A-5}\sum_{j=i+1}^{A}\sum_{k=i+1}^{A}\sum_{l=i+1}^{A}\sum_{m=k+1}^{A-1}\sum_{n=j+1}^{A} \dots\dots\sum_{A=s+1}^{A}(Xij.Xjk.Xkl.Xlm.Xmn.Xin +$$
$$Xni.Xji.Xkj.Xlk.Xml.Xnm)Zo \dots\dots Zs.ZA\ \boldsymbol{k,l,m,n} \dots\dots \boldsymbol{A \neq pus}$$

In general, the permanent of an M x M matrix, [A] with attributes x_{ij} defined by H Forbert et al. as:

$$\text{Per (X)} = \sum_{p}\prod_{i=1}^{N} Xi, P(i)$$

Where the sum is the overall permutations P. Keeping in mind the end goal to ascertain the improvement points, a PC program for the M x M framework was composed in C++ computer language. The program takes a couple of minutes to run the query and figure the estimation of the permanent function value. The legitimate stream graph for creating and structure of the program is connected in the appendix (A1). The evaluation of the permanent function is referred to as the "*Satisfaction Index*" for the proposed consider. A well-ordered way to deal with applying GTMA identifies as takes after the following procedure.

Step1: Selection of cases to be reviewed and factors against which their performance is considered, as mentioned above in discussions of Relative importance matrix.

Step2: Define the relative performance of each case concerning each other. The value of off-diagonal elements is referred from this selection for filling (Table 2).

TABLE 2
THE SCALE OF RELATIVE SATISFACTION FOR THE VALUES

Portraiture	**x_{ij} (Case A)**	**1-x_{ij} (Case B)**
Equally satisfied in both cases	0.5	0.5
One factor(i) is slightly more satisfactory than other(j)	0.6	0.4
One factor(i) is strongly satisfactory than other(j)	0.7	0.3
One factor(i) is very strongly satisfactory than other(j)	0.8	0.2
One factor(i) is extremely high in satisfaction while other is least satisfied(j)	0.9	0.1
One factor(i) is exceptionally high in satisfaction while other is least satisfied(j)	1.0	0.0

Step3: Find out the importance of each factor in observation for individual cases on the parameterized scale Gerhart (2005). This step derives the value of diagonal elements in the matrix (Table 3).

TABLE 3
THE SCALE OF RELATIVE SATISFACTION FOR THE VALUES

Qualitative Significance	**Scaled Value (Z_i)**
Exceptionally Bad	0.0
Extremely Bad	0.1
Very Bad	0.2
Bad	0.3
Below Average	0.4
Average	0.5
Above Average	0.6
Good	0.7
Very Good	0.8
Extremely Good	0.9
Exceptionally Good	1.0

Step4: The step requires developing the required matrix for individual cases. The MxM matrix developed is **Matrix [X]**. The values in the matrix are the result of the parametrized scale chosen in step2 and step3 for filling in the value of non-diagonal and diagonal elements.

Step5: Calculating the permanent function best applicable to individual cases in observation.

Step6: This step requires the use of calculated permanent function value for the individual cases, as illustrated in the above equation Per(X). The same can be achieved by using manual calculation; however, for higher accuracy and time constraint, a C^{++} program is used for the same. The compiler used is TurboC^{++}; however, another compiler like CodeBlocks can be used as well. This approach allows for a more straightforward comparison of each value to the different case values to determine the best performance in the cases.

Case Illustration

The sample case is based on the public survey, which was used to record various people as participants. The survey conducted through various online media sources ranging from non-finance services related to finance specialists. The survey was conducted to inspect the view of a variety of audiences ranging from industries specialist and ordinary people. The approach includes as many people as possible for acquiring distinct views and possible effects regarding observations with the changing backgrounds of the people, whether regarding their economic background, educational background, or professional background to better perceive the case in observation over the generalized market and economy.

The mode of information collection was chosen to be surveyed to facilitate the maximum amount of information collected over a smaller period with people of sparse backgrounds. The study conducted was in the form of a questionnaire with the no-costs to access links of Google Forms Link. The survey consisted of a total of 10 fundamental questions with 11 choices as a whole. The choices range from the Extremely Bad response to Extremely Good on the scale of 00 to 10 respectively and in-between. Each subject was further sub-branch for several cases to allow the participants to rank different parameters and their relative satisfaction in the presented cases. The cases mentioned are same as cases cited as alternatives in the methodology while each question parameter is same as the factor on which their relative performances are observed.

While the data collected in itself is allowably sorted to draw out the useful information correctly, the collected data is required to be further corrected to be able to provide more accurate information. The process can be solved via various methods to achieve higher accuracy, such as using median data values out of all participants; however, this type of approach may dilute the effectiveness in some cases. Similar results can be observed for other statistical data normalizations; thus it is profoundly prescribed to use an ample sample of the data to reduce the chances of high bias opinions and have a higher generalized point of view of the conventional markets. However, it is of advice that the sample space must always include people with mostly financial and managerial backgrounds or people having a basic understanding of economics and finances as a whole. This is to reduce the biased opinions out of the sample which may be shaped due to misinformation or other sources.

Here is an example is taken as the reference from many of participants in the survey to illustrate the basic idea of applying the discussed analytical method to find the best case out-performing relative to others on the selected specific factors. The expected result is suggested to be a significant value margin with maximum performance in the case of GST with the example of demonetization to be last with GST along with demonetization in the middle to near GST case values. The factors in observation can be reshuffled again and repeatedly for finding the best-suited elements to generate best and worst-performing factors combinations and sort-out the factors requiring maximum attention to examine for areas for possible implementations in future amendments.

STEP 1

This step entails the selection of the factors on which the various cases are being examined. The factors selected here are on the personal selection of the observer and can be varied upon the choice of another observer on the cases of the objective of the observer. The factors selected here are Ease of Filing (EoF), Government Incentives(GI), Foreign Trade Barriers(FTB), Total Monetary Payment(TMP), Customer Friendly(CF), Availability of Loans(AoL), Procurement of Materials(PoM), Supplier Management(SM), Warehousing(Wh) and Credit Cash Flow(CCF). A general summary of each alternative is provided in the summary table below for the review.
Also, along with the factors, the cases in observation selected are to be listed. The cases in selections here are:

Case I	Post-GST Period
Case II	Post-Demonetization Period
Case III	*Post-Demonetisation period with successively implemented GST* (Hypothetical)

STEP 2

Now the matrix X is filled with values of relative satisfaction based on the relative satisfaction index table. The values are selected for the off-diagonal elements by observer and survey results about the overall weight of each factor over the other. The values of diagonal elements are discussed ahead.

Matrix [X]

$[X]$	EOF	GI	FTB	TMB	CF	AoL	PoM	SM	Wh	CCF
EOF	$Z1$	0.5	0.6	0.3	0.5	0.6	0.8	0.7	0.7	0.4
GI	0.5	$Z2$	0.5	0.6	0.7	0.5	0.7	0.6	0.8	0.5
FTB	0.4	0.5	$Z3$	0.3	0.4	0.3	0.4	0.4	0.4	0.3
TMB	0.7	0.4	0.7	$Z4$	0.5	0.6	0.7	0.7	0.7	0.6
CF	0.5	0.3	0.6	0.5	$Z5$	0.4	0.4	0.4	0.4	0.3
AoL	0.4	0.5	0.7	0.4	0.6	$Z6$	0.6	0.4	0.4	0.5
PoM	0.2	0.3	0.6	0.3	0.6	0.4	$Z7$	0.5	0.6	0.4
SM	0.3	0.4	0.6	0.3	0.6	0.4	0.5	$Z8$	0.4	0.4
Wh	0.3	0.2	0.6	0.3	0.6	0.4	0.4	0.6	$Z9$	0.4
CCF	0.6	0.5	0.7	0.4	0.7	0.5	0.6	0.6	0.6	$Z10$

STEP 3

The values for the diagonal elements selected from the set scale mentioned are in the methodology. The values selected by the survey and participants' choices for the particular cases in each (Table 4 and Figure 2).

TABLE 4
VALUES OF DIAGONAL ELEMENTS IN THE INDIVIDUAL CASES

Factors	Case I	Case II	Case III
EoF	0.8	0.3	0.8
GI	0.8	0.1	0.6

FTB	0.8	0.5	0.6
TMP	0.8	0.3	0.5
CF	0.7	0.2	0.5
AoL	0.6	0.1	0.1
PoM	0.8	0.1	0.5
SM	0.7	0.4	0.5
Wh	0.8	0.3	0.6
CCF	0.9	0.4	0.7

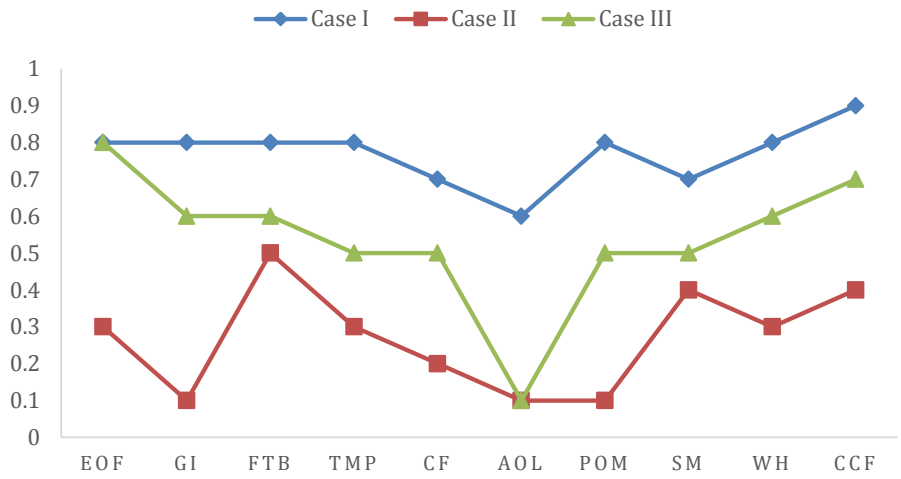


FIGURE 2
RELATIVE SATISFACTION OF EACH FACTOR IN EACH CASE

STEP 4

This step requires the fulfillment of the matrix values as required for the individual cases in the diagonal of the matrix developed in STEP 2. The values for Case I, Case II, and Case III is taken from the table developed in STEP 3.

Case I Matrix

$[X]$	EOF	GI	FTB	TMB	CF	AoL	PoM	SM	Wh	CCF
EOF	0.8	0.5	0.6	0.3	0.5	0.6	0.8	0.7	0.7	0.4
GI	0.5	0.8	0.5	0.6	0.7	0.5	0.7	0.6	0.8	0.5
FTB	0.4	0.5	0.8	0.3	0.4	0.3	0.4	0.4	0.4	0.3
TMB	0.7	0.4	0.7	0.8	0.5	0.6	0.7	0.7	0.7	0.6
CF	0.5	0.3	0.6	0.5	0.7	0.4	0.4	0.4	0.4	0.3
AoL	0.4	0.5	0.7	0.4	0.6	0.6	0.6	0.4	0.4	0.5
PoM	0.2	0.3	0.6	0.3	0.6	0.4	0.8	0.5	0.6	0.4
SM	0.3	0.4	0.6	0.3	0.6	0.4	0.5	0.7	0.4	0.4
Wh	0.3	0.2	0.6	0.3	0.6	0.4	0.4	0.6	0.8	0.4
CCF	0.6	0.5	0.7	0.4	0.7	0.5	0.6	0.6	0.6	0.9

Case II Matrix

$[X]$	EOF	GI	FTB	TMB	CF	AoL	PoM	SM	Wh	CCF
EOF	0.3	0.5	0.6	0.3	0.5	0.6	0.8	0.7	0.7	0.4
GI	0.5	0.1	0.5	0.6	0.7	0.5	0.7	0.6	0.8	0.5
FTB	0.4	0.5	0.5	0.3	0.4	0.3	0.4	0.4	0.4	0.3
TMB	0.7	0.4	0.7	0.3	0.5	0.6	0.7	0.7	0.7	0.6
CF	0.5	0.3	0.6	0.5	0.2	0.4	0.4	0.4	0.4	0.3
AoL	0.4	0.5	0.7	0.4	0.6	0.1	0.6	0.4	0.4	0.5
PoM	0.2	0.3	0.6	0.3	0.6	0.4	0.1	0.5	0.6	0.4
SM	0.3	0.4	0.6	0.3	0.6	0.4	0.5	0.4	0.4	0.4
Wh	0.3	0.2	0.6	0.3	0.6	0.4	0.4	0.6	0.3	0.4
CCF	0.6	0.5	0.7	0.4	0.7	0.5	0.6	0.6	0.6	0.4

Case III Matrix

$[X]$	EOF	GI	FTB	TMB	CF	AoL	PoM	SM	Wh	CCF
EOF	0.8	0.5	0.6	0.3	0.5	0.6	0.8	0.7	0.7	0.4
GI	0.5	0.6	0.5	0.6	0.7	0.5	0.7	0.6	0.8	0.5
FTB	0.4	0.5	0.6	0.3	0.4	0.3	0.4	0.4	0.4	0.3
TMB	0.7	0.4	0.7	0.5	0.5	0.6	0.7	0.7	0.7	0.6
CF	0.5	0.3	0.6	0.5	0.5	0.4	0.4	0.4	0.4	0.3
AoL	0.4	0.5	0.7	0.4	0.6	0.1	0.6	0.4	0.4	0.5
PoM	0.2	0.3	0.6	0.3	0.6	0.4	0.5	0.5	0.6	0.4
SM	0.3	0.4	0.6	0.3	0.6	0.4	0.5	0.5	0.4	0.4
Wh	0.3	0.2	0.6	0.3	0.6	0.4	0.4	0.6	0.6	0.4
CCF	0.6	0.5	0.7	0.4	0.7	0.5	0.6	0.6	0.6	0.7

STEP 5

The permanent function can be computed for each matrix, as described in the methodology. The value of the permanent function can be calculated using the Computer software programs as employed here. The value is computed using a C++ program in this research.

STEP 6

The Value of each case is calculated using the desired techniques and tools. The values here are arranged in the order of the cases computed.

CASE	*Permanent Function Computed Value*
Case I	4572.81
Case II	1626.06
Case III	3156.91

The values of the permanent function suggest a simple trend. The value of the permanent function in Case I is indicative of a high level of general satisfaction for the given selected factors in observation. These values correlate with the initial hypothesis that the new GST tax regime out to have a higher satisfaction index in overall comparison. The value of Case II is significantly lower as compared to the other cases, which also adds up to the initial hypothesis. The execution of the demonetization also received backlash for the way of executions and other types of after-effects that lingered in the commercial businesses for a more extended period. Case III was an altered hypothetical case in which the consideration was taken for if the demonetization had been implemented after the assimilation of GST in previous tax-regime structure. This case shows a significantly closer value to Case I, where only GST is implemented in typical routine markets. This value is significantly better than the case of application of demonetization in the previous ta regimes and thus clearly suggests the merits of GST and its robustness to allow for cushioning such unexpected blows to the current Indian economy (Figure 3).

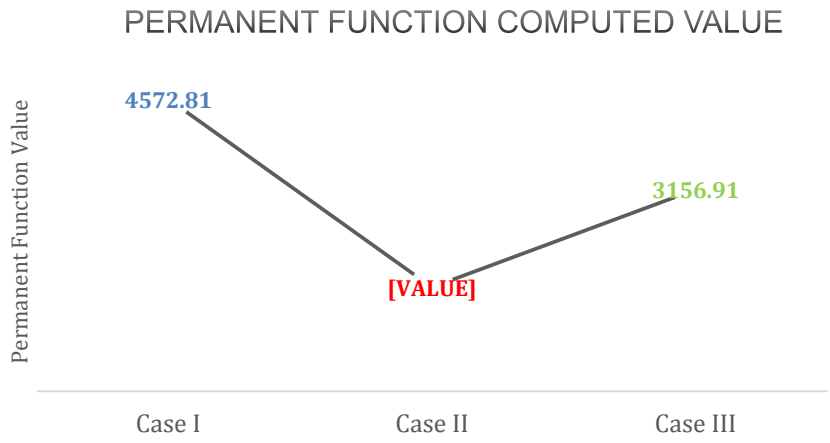


FIGURE 3
COMPARATIVE TREND LINE FOR VALUES OF PERMANENT FUNCTION IN EACH CASE

CONCLUSION

The study undertaken in this paper uses the technique of GTMA (Graph Theory and Matrix Approach) for conducting a study about GST and its characteristics by individual factors which are variable by the type of study propose and the nature of observation required. The factors selected here Ease of Filing (EoF), Government Incentives (GI), Foreign Trade Barriers (FTB), Total Monetary Payment (TMP), Customer Friendly (CF), Availability of Loans (AoL), Procurement of Materials (PoM), Supplier Management (S), Warehousing (Wh) and Credit Cash Flow(CCF) are selected on the discretion of the observer as the motive of study was to find out about the collective satisfaction of people before and after GST implementation in India. The main reason for the selection of the GTMA technique is because of the ease of application, and

the results generated are satisfactorily accurate. The highlight of the approach is the permanent function, employed to find an overall comparative value of each case.

Here the value generated by the permanent function is taken as "*Overall Satisfaction Index*." The cases can be reiterated with different sets of factors in play and with different cases in question. All values generated and thus can be compared to find the places of improvements within the set of factors. This type of study is substantially free from constraints of the factors and cases in observation and allows for selecting the best cases along with better factors for application of the particular ideas in any the situation with the best possible outcome. The case discussed here is a culmination of data collected via means of an online survey. The results generated here are in support of the hypothesis of an observer than GST is better for upcoming Indian economy. The observations here can also help for amendments of the current GST status in the economy for broader and better fetching results. The technique used here is applicable for different types of applications as well and is not supposed to be specific to these types of analysis. Also, the results generated here can only be taken as a model representation of these types of analysis required a more substantial sample survey that may be employed here.

Acknowledgments

We much acknowledge the contributions of the subject editor, language expert, and reviewers for helping us in improving the quality of this article.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

Agrawal, S., Singh, R. K., & Murtaza, Q. (2016). Disposition decisions in reverse logistics: Graph theory and matrix approach. *Journal of Cleaner Production*, 137, 93-104.

Dani, S. (2016). A Research Paper on an Impact of Goods and Service Tax (GST) on Indian Economy. *Bus Eco* J, 7, 264.

Deloitte GST Report retrieved from https://www2.deloitte.com/content/dam/Deloitte/in/Documents/tax/in-tax-gst-in-india- taking-stock-noexp.pdf.

Forbert, H., & Marx, D. (2003). Calculation of the permanent of a sparse positive matrix. Computer Physics Communications, 150(3), 267-273.

Gerhart, B. (2005). The (affective) dispositional approach to job satisfaction: Sorting out the policy implications. *Journal of Organizational Behavior*, *26*(1), 79-97.

Goods and Service Tax. Retrieved from https://www.gst.gov.in/about/gst/history.

GST has not affected prices, but increased compliance burden: KPMG survey Retrieved from http://www.business-standard.com/article/economy-policy/gst-has-not-affected-prices-but-increased-compliance-kpmg-survey-118022300762_1.html.

Jain, V., & Raj, T. (2016). Modeling and analysis of FMS performance variables by ISM, SEM and GTMA approach. *International Journal of Production Economics*, 171, 84-96.

Malik, S., Kumari, A., & Agrawal, S. (2015). Selection of locations of collection centers for reverse logistics using GTMA. Materials Today: *Proceedings, 2*(4-5), 2538-2547.

Moorman, R.H., Niehoff, B.P., & Organ, D.W. (1993). Treating employees fairly and organizational citizenship behavior: Sorting the effects of job satisfaction, organizational commitment, and procedural justice. *Employee responsibilities and rights journal, 6*(3), 209-225.

Muduli, K., Govindan, K., Barve, A., Kannan, D., & Geng, Y. (2013). Role of behavioral factors in green supply chain management implementation in Indian mining industries. Resources, Conservation and Recycling, 76, 50-60.

Rao, R.V. (2007). Decision making in the manufacturing environment: using graph theory and fuzzy multiple attribute decision making methods. Springer Science & Business Media.

Rao, R.V., & Padmanabhan, K. K. (2007). Rapid prototyping process selection using graph theory and matrix approach. *Journal of Materials Processing Technology, 194*(1-3), 81-88.

Richard, P.J., Devinney, T. M., Yip, G. S., & Johnson, G. (2009). Measuring organizational performance: Towards methodological best practice. *Journal of management, 35*(3), 718-804.

Slack, N. (1994). The importance-performance matrix as a determinant of improvement priority. International *Journal of Operations & Production Management, 14*(5), 59-75.

Toloo, M., & Nalchigar, S. (2011). A new DEA method for supplier selection in presence of both cardinal and ordinal data. *Expert Systems with Applications, 38*(12), 14726-14731.

Venkatesh, A., Vulugundam, A. (2015). Impact on Of Gst on Supply Chain Strategy and Its Effect on Warehousing And Transportation http://www.mahindralogistics.com/logiquest/assets/pdf/LogiQuest_EastCoastExpress_NMIMS.PDF

Wagner, S.M., & Neshat, N. (2010). Assessing the vulnerability of supply chains using graph theory. *International Journal of Production Economics, 126*(1), 121-129.

Yang, J.L., & Tzeng, G.H. (2011). An integrated MCDM technique combined with DEMATEL for a novel cluster-weighted with ANP method. *Expert Systems with Applications, 38*(3), 1417-1424.

Hiding Data in Images Using Cryptography and Deep Neural Network

Kartik Sharma[1, *], Ashutosh Aggarwal[1], Tanay Singhania[2], Deepak Gupta[1], Ashish Khanna[1]

[1] Computer Science and Engineering Department, Guru Gobind Singh Indraprastha University, New Delhi, India
Email: 98kartik.sharma@gmail.com; ashutoshaggarwal98@gmail.com; deepakgupta@mait.ac.in; ashishkhanna@mait.ac.in
[2] Applied Mathematics Department, Delhi Technological University, New Delhi, India
Email: tanay_bt2k17@dtu.ac.in
*Corresponding Author: Kartik Sharma, Email: 98kartik.sharma@gmail.com

Abstract

Steganography is an art of obscuring data inside another quotidian file of similar or varying types. Hiding data has always been of significant importance to digital forensics. Previously, steganography has been combined with cryptography and neural networks separately. Whereas, this research combines steganography, cryptography with the neural networks all together to hide an image inside another container image of the larger or same size. Although the cryptographic technique used is quite simple, but is effective when convoluted with deep neural nets. Other steganography techniques involve hiding data efficiently, but in a uniform pattern which makes it less secure. This method targets both the challenges and make data hiding secure and non-uniform.

Keywords

Image Steganography, Cryptography, Convolutional Neural Network, Deep Learning, Digital Data Security

1. Introduction

The ingenuity of enshrouding a file within a different file is called Steganography that has been in exercise from the past 2500 years. Better steganography definitions are given by Provos and Honeyman [1]; Johnson and Jajodia [2]; Morkel et al. [3]; Fridrich and Goljan [4]. Predominantly, in steganography, there is a secret message and a carrier and the arcane message is hidden in the carrier in some way that it is undetectable. Therefore, it has multiple applications in defense organizations, intelligence agencies, identity cards, etc.

Steganography is mainly of 4 types; image, text, audio/video and protocol which are well explained by Shih [5] and Morkel et al. [3].

The method used to recover the message hidden by steganography is called Steganalysis. Steganalysis is lucidly explained by Shih [5]; Fridrich and Goljan [4]; Avcibas et al. [6]. Hitherto, steganography was effortless to implement securely as people did not know much about it. But with the headway in the Steganalysis, encrypting the secret message before passing it to a carrier has become a prerequisite. This paper mainly focuses on Image steganography as it is one of the most vastly used techniques.

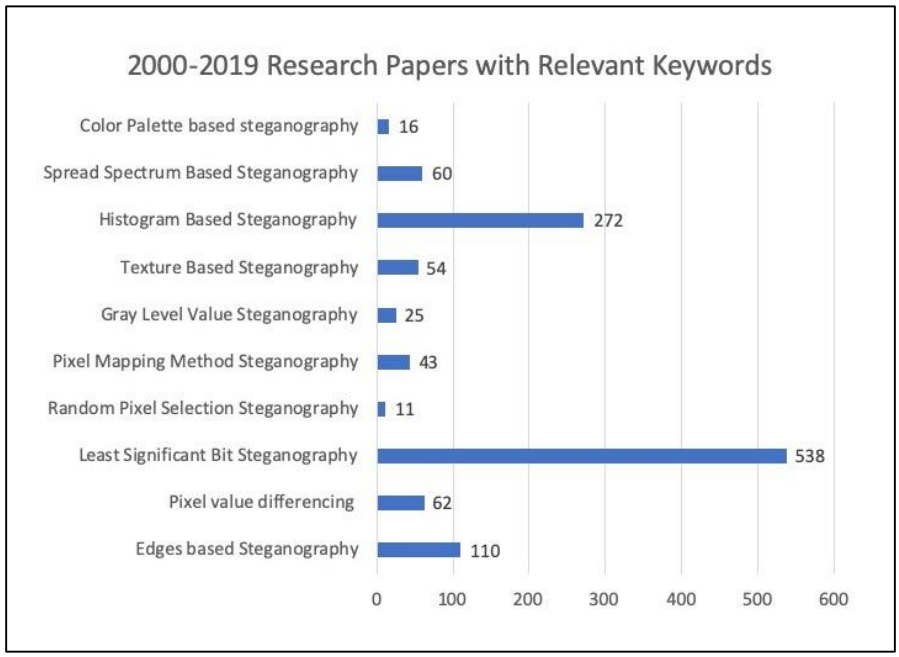


Figure 1. Research Papers (2000-2019) with relevant Keywords.

Now, there are profuse techniques for Image Steganography as explained by Li et al. [7]; Subhedar et al. [8]. The bar graph plotted in Figure 1 exhibits the frequency of different techniques used. Out of these techniques, the most common method is the Least Significant Bit Steganography. Therefore, this habitual and widely accepted method is used. Moreover, the Least Significant Bit Cryptography easily dovetails with Deep Neural Networks, which is explained further in this paper.

In LSB Steganography, data is hidden in the least significant bit of the carrier [9], [10]. While talking about images in specific, alteration is performed on the last bit from the 8-bit carrier pixel value. Due to changing only the least significant bits of the cover image, it goes unnoticed to the naked eye that there is a slight change in the color values. A few years back, it was a good practice, but as the method disseminated, Steganalysts found a method to recover it, which made it no longer usable. Secret image can be easily found by extracting the least significant bits of the

cover image.

Different tools such as ZSteg, JSteg, StegoVeritas, etcetera. can be used to find the Least Significant bit values. As using the LSB technique became trite, Baluja [11] implemented its variant with Deep Neural Networks to make it secure. The method proposed by Baluja [11], had a constraint that the image could not be made public. Addressing this specific constraint, an improved algorithm involving cryptography and deep neural networks is discussed in the paper.

Theoretical, as well as the practical implementation of the algorithm, is discussed in the paper. The algorithm is trained on a public dataset, producing an accuracy of 90% approximately on both training and test set, therefore making it an effective algorithm for data hiding.

The paper proposes the following:

- An improved algorithm for Steganography using Cryptography and Deep Neural Networks is propounded for hiding data.
- This algorithm uses Convolutional Neural Networks with Adam optimizer for both hiding and revealing part of the architecture.
- Public Dataset is used for training this algorithm, making the accuracy of approximately 90% on both train and test dataset.

The remaining paper is structured as follows. The second section includes reviews of varying papers. The third and fourth section contains the methodology and results respectively. Section 5 explains the advantages of using this algorithm. Section 6 and 7 describe conclusion and future work respectively.

2. Literature Review

The paper propounded by Morkel et al. [3] focused mainly on Image steganography while briefly explaining other steganography techniques. The paper proposed various techniques to hide data in images and a detailed comparison of various steganography techniques.

Li et al. [7] provided a survey on Steganalysis and steganography for images. They mainly used JPEG format images for their research and provided extensive research explaining different steganographic techniques as well as the attack scenarios on those methods. Different tools such as OutGuess, JSteg/JHide, and methodologies such as LSB, LSB Matching, Stochastic steganography, etc. were used for demonstration.

Anderson et al. [12] explained various limitations in image-based steganography when using these theoretical concepts privately. Therefore, they proposed public-key steganography with parity checks to amplify covertness. Public key steganography was demonstrated using passive and sometimes with the active warden as well.

Marwaha and Marwaha [13] used Asymmetric Key Cryptography to encrypt the

message. Specifically, the Data Encryption Standard (DES) was used to encrypt the data. Furthermore, the cipher was hidden in the image file. This technique worked only when the container image was transmitted in a lossless environment. Data Encryption Algorithm (DEA) was used to extract the secret text back from the image.

Baluja [11] implemented the Least Significant Bit steganography using convolutional neural networks. He was successfully able to hide and reveal the image. However, his research has a downside that if the cover image is made public, the secret image could be extracted from the cover by amplifying the difference.

Research presented by Djebbar et al. [14] showcased various audio steganography techniques. Different methods such as low bit encoding, magnitude spectrum, wavelet coefficients, bitstream hiding, etc. were evaluated in several hiding domains such as temporal, frequency, etc. They compared the methods mentioned earlier on the bases of hiding rate, capacity available, distortions, robustness, and lossy compression.

Biometric data protection was well explained by Douglas et al. [15]. They explained various biometric techniques such as fingerprint, retina, face and their issues. These different biometric were further used to hide data using some well-known image steganography methods like Spatial Domain Techniques (LSB-Steganography), Pixel Value Differencing (PVD), Transform Domain Techniques (DCT/ DWT) and some hybrid techniques such as Singular Value Decomposition (SVD). Binarization was incorporated to hide data inside biometric data.

Jiang et al. [16] proposed a new LSB based Quantum Steganography Algorithm for images. The two algorithms presented were advantageous as they as they were absolutely blind as well as no classical computers were required to perform this. Image preprocessing was done using the Hilbert image scrambling algorithm that permuted the pixels into new positions.

Wang et al. [17] briefly explained various steganography techniques and tools. Different steganalysis techniques to break the above mentioned steganographic methodologies such as POV-based Chi-Square test, Palette Checking, Universal Blind Detection, etc. were succinctly expounded.

3. Methodology

3.1. Artificial Neural Network

Artificial Neural Network (ANN) is a progression of an algorithm that behaves like a human sensory system [18]. An ANN embodies a huge number of interconnected processing units called nodes (neurons) that work collaboratively to solve a fastidious errand. In a similar fashion to our brain, the nodes work in parallel and each node collects input from others. The nodes then compute these inputs and pass on the data to the next connected node. Neural networks learn to perform tasks by analyzing pre-

defined data.

Convolutional Neural Network (CNN) is one of the widely used algorithms for Neural Networks. CNN or ConvNet is mostly used to analyze images. CNN's allow us to encode image-specific features into the architecture, making the network more fitting for the tasks, handling the images - whilst reducing the parameters required to set up the model. A better understanding can be found in [19], [20].

3.1.1. Convolutional Neural Network

CNNs are very powerful neural networks that can be classified as a regularized version of Multilayer Perceptrons or MLP [21]. The name “convolutional neural network” indicates that the network employs a mathematical operation called convolution [22]. Convolution is a specific kind of linear operation that expresses the amount of overlap of one function as it is shifted over another function. Convolutional networks are simply neural networks that use convolution in place of general matrix multiplication in at least one of their layers [23]-[25]. Reported studies [26]-[28] suggest that CNN based architecture for encoding and decoding features have the following advantages:

- CNN extract features of the images automatically.
- CNN effectively uses adjacent pixel information to effectively downsample the image first by convolution and then uses a prediction layer at the end.
- CNN performs well and gives better accuracy.

Using a deep neural network, CNN, in this case, one will be able to have a good sense of the patterns of natural images. The network will be able to make judgments on which area is redundant, and more pixels can be hidden there. By conserving the space on unnecessary areas, the amount of hidden data can be increased. Since the structure and the weights can be randomized, the network will hide the data that cannot be known to anybody who doesn't have the weights [29], [30].

3.2. Architecture

The network architecture is somewhat similar to Auto-Encoders [31]. In general, Auto-Encoders are used to reproduce the input ensuing a set of transformations. By performing this, they learn about the characteristics of the input distribution.

But in our illustration, the proposed architecture is marginally distinct. Rather than simply procreating images, the network has to conceal an image, additionally procreate another image. The proposed architecture is exhibited in Figure 2.

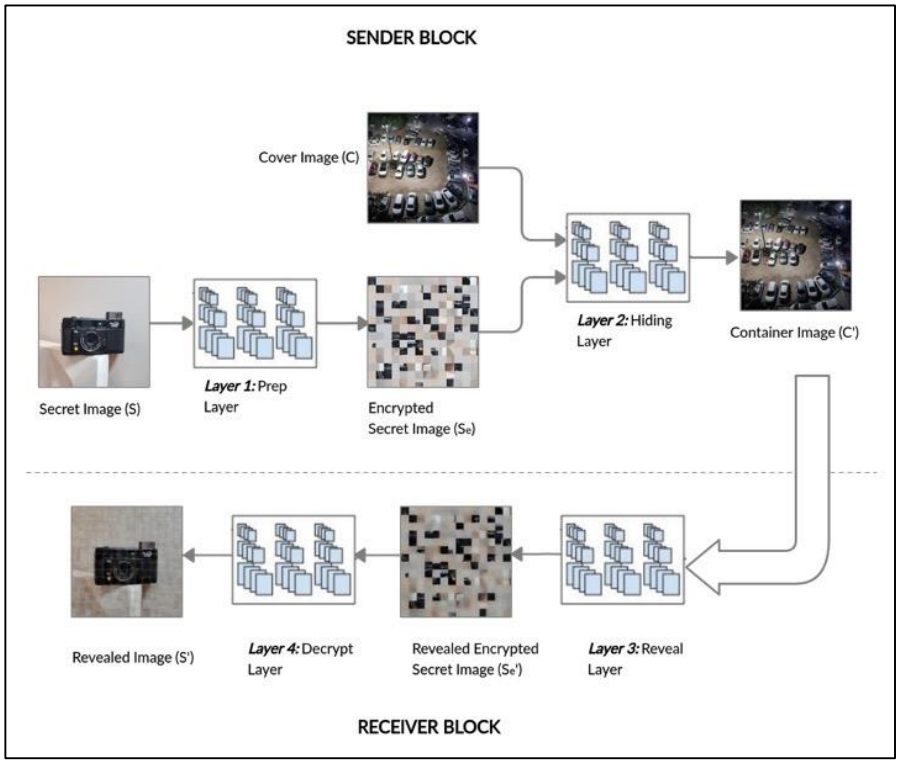


Figure 2. Architecture of the proposed system. Top: Sender Block. Bottom: Receiver Block

The four components shown in Figure 2 are trained as a single network but it will be easier to explain them individually. The first layer, Prep Layer, prepares the secret image to be hidden. This layer serves various purposes. Firstly, if the secret image is smaller than the cover image, it increases the size of the secret image to the size of the cover image, hence distributing the secret image's bits across the entire pixels. Furthermore, relevant to all measurements of hidden images, the purpose is to reconstruct the color-based pixels to more useful features for succinctly encoding the image such as edges. The most important purpose of this layer was to finally embed the secret image to the encrypted image so as to avoid any leak of the secret message.

The second layer, Hiding Layer, takes the output of the Prep Layer as input and a cover image so as to produce the Container Image. The input to this network is a square pixel image, with depth considered of RGB channels of the cover image and the reconstructed channels of the secret image. These two layers together form the Sender Block. The Container image produced can be shared with the receiver.

The third Layer, Reveal Layer, is used by the receiver to produce back the encrypted image. This layer takes in the Container image as input and removes the cover image to generate the encrypted secret image.

The fourth Layer, Decrypt Layer, takes in the output of the Reveal Layer and decrypt the image to finally reveal the secret image. The third and fourth layer together forms the Receiver Block.

As shown by Baluja [11], the loss in the network can be computed by the given formula:

$$L(c,c',s,s') = \|c - c'\| + b\|s - s'\| \quad (1)$$

where the variables are as mentioned in Figure 2. The loss is the conventional Mean Square Error between the original cover image and the container image, and β*(MSE between the original secret image and the revealed image). β is a hyper-parameter that regulates how much the secret image should be restored. Since the given function is differentiable, the neural network can be trained end-to-end.

3.3. Optimizer

The role of the optimizer is to improve the weight parameters to minimize the loss function. It leads the loss function to find its global minima. There are various types of optimizers, for example - Momentum, Nesterov Accelerated Gradient, Adagrad, Adadelta, Adam, etc.

The model uses ADAM - Adaptive Moment Estimation [32] optimizer that calculates a separate learning rate for each parameter. It is computationally productive and has very low memory requirements which make it optimum for our model. The reason behind not using Momentum, Stochastic Gradient Descent, and Nestorov optimizers is because the dataset used is sparse. Other adaptive learning methods such as RMSprop, Adagrad, and Adadelta were outperformed by the ADAM optimizer making it the ideal choice.

ADAM computes the mean and uncentered variance of past gradients as p_t and q_t with β1, β2 ∈ [0, 1) as hyper-parameters respectively.

$$p_t = b_1 p_t - 1 + (1 - b_1) g_t \quad (2)$$

$$q_t = b_2 p_t - 1 + (1 - b_2) g^2{}_t \quad (3)$$

Further, it computes bias-corrected first moment (p_t) and second moment (q_t) as follows:

$$\hat{p}_t = \frac{p_t}{1 - b_1^t} \quad (4)$$

$$\hat{q}_t = \frac{q_t}{1 - b_2^t} \quad (5)$$

Then these bias-corrected mean and variance of past gradients are used to update

the parameters of the model as follows:

$$q_{t+1} = q_t - \frac{n}{\sqrt{\hat{q}_t + e}} \hat{p}_t \tag{6}$$

where ε is a small scalar used to prevent division by zero and g_t refers to the loss function of the model.

3.4 Image Encryption

Images are widely used as part of communication since they are more effortless to process than text. As explained by Kumar et al. [33], the user might communicate on a network that is compromised and security of the information becomes crucial in such cases. Hence, encryption or the scheme used to transform an image into another image that is not easily comprehensible [34], [35] is important to maintain the security of the information being communicated. Image encryption has applications in various areas - for example, military, security agencies or wherever the data is sensitive or contains privileged information.

Image Encryption methods or algorithms revolve around the following three ideas:

- Pixel Permutation - Scrambling the pixels
- Pixel Substitution - Modifying each pixel value
- Visual Alteration

An image histogram is one of the security parameters kept in consideration for encryption methods as explained by Li et al. [36]. The histogram displays the frequency distribution and provides an insight into the frequency of each pixel value. In the various encryption methods as explained by Khan and Shah [37], methods based on the transformation of pixel values for example - AES (Advanced Encryption System) tend to create a uniform image histogram that protects it from plain text attack but requires zero loss while decrypting the image. This is not useful in this case as the neural network used (to hide images in other images) depends on the redundant area in the cover images which in turn gives the amount of hidden data. Similarly, ECC (Elliptic Curve Cryptography) as explained by Geetha et al. [38] cannot be used in our scenario.

Due to the data loss in the neural networks, methods like AES and ECC fail. Hence, one can use methods based on scrambling of data (in this case pixels) retaining the same pixel value but at a different location giving the same histogram [39]. The scrambling helps in providing the same histogram but decreases the correlation between pixels.

In this paper, scrambling or shuffling blocks of pixels technique is used [40]. As discussed by Kumar et al. [33], displacement techniques give the same image

histogram but a reduced correlation between adjacent pixels of the image as illustrated in Figure 3.

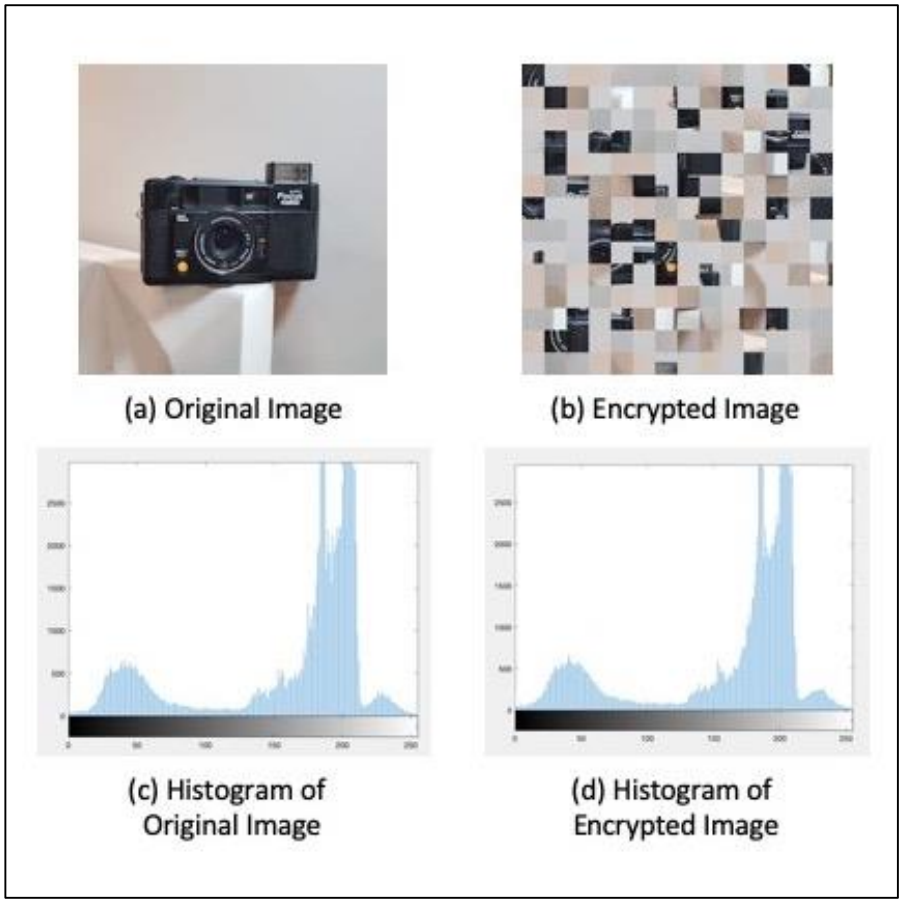


Figure 3. Histogram comparison of the original and encrypted image

As the number of blocks increases, the correlation decreases giving a suitable encryption layer. The number of blocks can be referred to as the order of encryption. By increasing the order of encryption it is observed that one cannot interpret the significance of the image just by looking at it. Figure 4 depicts the effect of the increasing order of encryption.

After experimenting, it is observed that the order of encryption 196 gives the most suitable level of encryption and low correlation. Moreover, 196 blocks of pixels mean 196! ways to arrange them. This gives ≅ 5.08 x 10365 number of permutations and considering a computer does 1016 calculations per second then it will take about ≅ 10342 years to find the right permutation. Hence it is used as the first layer of encryption in the network.

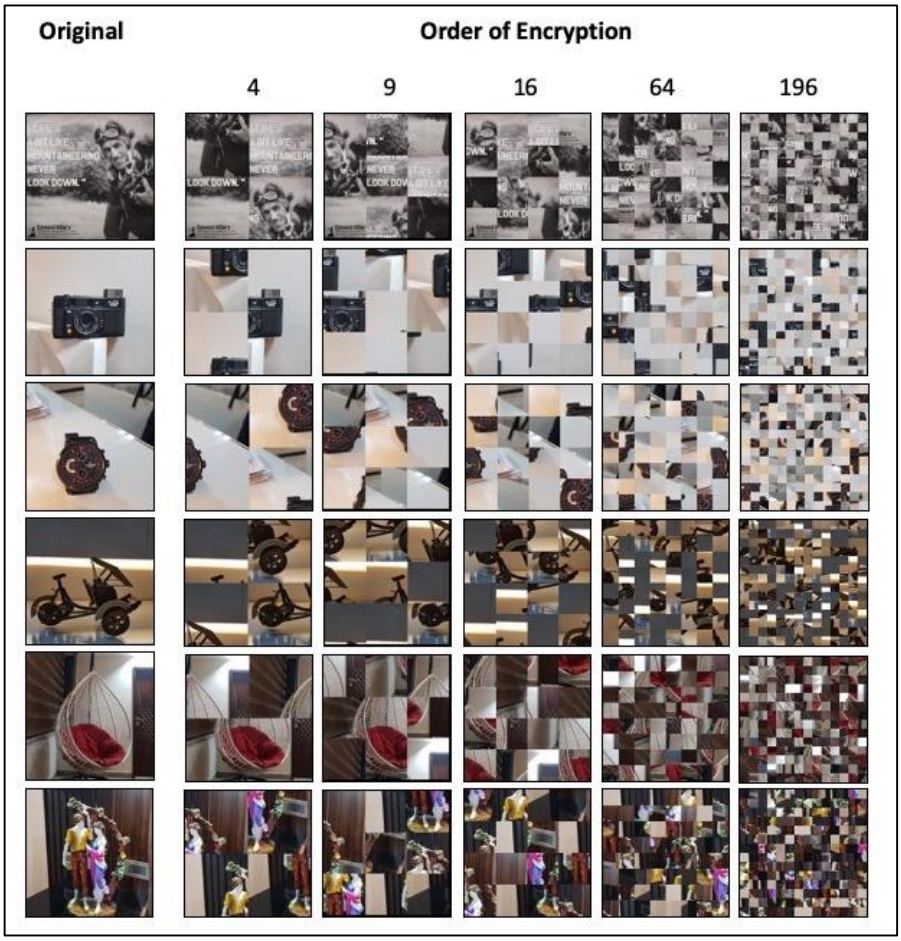


Figure 4. Encryption order (increasing left to right) of secret image comparison chart

3.5. Integrating Neural Networks with Encryption

3.5.1 Experimental Configuration

ML-Engine from Google Cloud Platform is used to train the model. JupyterLab on AI Platform's Notebook Instances which is currently in the β phase is used. In particular, we used Tensorflow 2.0 Notebook Instance that has pre-installed Tensorflow with Keras support. It also had all other essential artificial intelligence and machine learning libraries compiled in it. Model was developed using libraries including Tensorflow, Keras, Matplotlib, Numpy, Scipy, etc.

Table 1. Software and Hardware Specifications

COMPONENT	SPECIFICATIONS
GPU	1 NVIDIA Tesla K80 having 2496 CUDA(version 10.0) Cores, 12GB(11.46 Usable) DDR5 VRAM
CPU	4 Intel Xeon CPU @ 2.20GHz
RAM	15GB
Boot disk	100 GB storage space available
Language	Python 3.5.3
OS	Debian GNU/Linux 9.11

3.5.2. Dataset Preparation and Encryption

The Flickr30k dataset has been used to train the model. The images were of irregular size, therefore were scaled at 256x256 as per our training model. A total of 31,783 images were used which had an RGB scale. Datasets can be found at [41]. Furthermore, each image of the dataset is transformed into blocks that are scrambled thereafter.

3.5.3. Train Neural Net

To train the convolutional neural network, the B values being 0.25, 0.75 and 1 as represented in Table 5 have been used. A batch size of 32 images with 1000 epochs were used to train the model. Neural network with the above mentioned hardware took about 8 hours to train for a particular beta value. We used 3 nodes (3x3, 4x4 and 5x5) in every layer. 2 nodes (conv_prep0 and conv_prep1) were used in the preparation network. Whereas 5 layers (conv_hide0, conv_hide1, conv_hide2, conv_hide3, conv_hide4) were used in hiding network as represented in Table 3. For reveal network, 5 layers (conv_rev0, conv_rev1, conv_rev2, conv_rev3, conv_rev4) as well which were integrated with Adam optimizer were used as shown in Table 4. The following layers are summarized briefly in Table 2. Retrieving the encrypted image can be revealed using the decryption methodology explained under the heading Image Encryption.

Table 2: Summary of the proposed neural network

Model: “model_1”			
Layer(type)	Output Shape	Param #	Connected to
input_1 (InputLayer)	(None, 256, 256, 3)	0	
input_2 (InputLayer)	(None, 256, 256, 3)	0	
Encoder (Model)	(None, 256, 256, 3)	293273	input_1[0][0] input_2[0][0]
Decoder (Model)	(None, 256, 256, 3)	195388	Encoder[1][0]
concatenate_14 (Concatenate)	(None, 256, 256, 6)	0	Decoder[1][0] Encoder[1][0]
Total params: 488,661 Trainable params: 293,273 Non-trainable params: 195,388			

Table 3: Detailed representation of the Encoder model

Model: “Encoder”		
Layer(type)	Output Shape	Connected to
input_3	(None, 256, 256, 3)	
conv_prep0_(3x3 4x4 and 5x5)	(None, 256, 256, (50, 10, 5))	input_3[0][0]
concatenate_1	(None, 256, 256, 65)	conv_prep0_(3x3, 4x4, 5x5)[0][0]
conv_prep1_(3x3, 4x4 and 5x5)	(None, 256, 256, (50, 10, 5))	concatenate_1[0][0]
input_4	(None, 256, 256, 3)	
concatenate_2	(None, 256, 256, 65)	conv_prep1_(3x3, 4x4 and 5x5)[0][0]
concatenate_3	(None, 256, 256, 68)	input_4[0][0], concatenate_2[0][0]
conv_hid0_(3x3, 4x4, 5x5)	(None, 256, 256, (50, 10, 5))	concatenate_3[0][0]
concatenate_4	(None, 256, 256, 65)	conv_hid0_(3x3, 4x4,

		5x5)[0][0]
conv_hid1_(3x3, 4x4, 5x5)	(None, 256, 256, (50, 10, 5))	concatenate_4[0][0]
concatenate_5	(None, 256, 256, 65)	conv_hid1_(3x3, 4x4, 5x5)[0][0]
conv_hid2_(3x3, 4x4, 5x5)	(None, 256, 256, (50, 10, 5))	concatenate_5[0][0]
concatenate_6	(None, 256, 256, 65)	conv_hid2_(3x3, 4x4, 5x5)[0][0]
conv_hid3_(3x3, 4x4, 5x5)	(None, 256, 256, (50, 10, 5))	concatenate_6[0][0]
concatenate_7	(None, 256, 256, 65)	conv_hid3_(3x3, 4x4, 5x5)[0][0]
conv_hid4_(3x3, 4x4, 5x5)	(None, 256, 256, (50, 10, 5))	concatenate_7[0][0]
concatenate_8	(None, 256, 256, 65)	conv_hid4_(3x3, 4x4, 5x5)[0][0]
output_C (Conv2D)	(None, 256, 256, 3)	concatenate_8[0][0]
Total params: 293,273 Trainable params: 293,273 Non-trainable params: 0		

Table 4: Detailed representation of the Decoder model

Model: “Decoder”		
Layer(type)	Output Shape	Connected to
input_5 (Input Layer)	(None, 256, 256, 3)	
output_C_noise (GaussianNoise)	(None, 256, 256, 3)	input_5[0][0]
conv_rev0_(3x3, 4x4 and 5x5)	(None, 256, 256, (50,10,5)	output_C_noise[0][0]
concatentate_9	(None, 256, 256, 65)	conv_rev0_(3x3, 4x4 and 5x5)[0][0]
conv_rev1_(3x3, 4x4 and 5x5)	(None, 256, 256, (50,10,5)	concatenate_9[0][0]
concatentate_10	(None, 256, 256, 65)	conv_rev1_(3x3, 4x4 and 5x5)[0][0]

conv_rev2_(3x3, 4x4 and 5x5)	(None, 256, 256, (50,10,5)	concatenate_10[0][0]
concatentate_11	(None, 256, 256, 65)	conv_rev2_(3x3, 4x4 and 5x5)[0][0]
conv_rev3_(3x3, 4x4 and 5x5)	(None, 256, 256, (50,10,5)	concatenate_11[0][0]
concatentate_12	(None, 256, 256, 65)	conv_rev3_(3x3, 4x4 and 5x5)[0][0]
conv_rev4_(3x3, 4x4 and 5x5)	(None, 256, 256, (50,10,5)	concatenate_12[0][0]
concatentate_13	(None, 256, 256, 65)	conv_rev4_(3x3, 4x4 and 5x5)[0][0]
output_S (Conv2D)	(None, 256, 256, 3)	concatenate_13[0][0]
Total params: 390,776 Trainable params: 195,388 Non-trainable params: 195,388		

4. Results

Table 5: MSE per pixel each for cover and secret.

β	COVER	SECRET
1	13.8137	11.2884
0.75	12.1389	13.6526
0.25	8.81926	17.5474

Table 5 represents the mean of Sum of Squared Errors per pixel each for cover and secret images. According to which β = 1 shows the minimum error in the secret image which is beneficial since data lost in-network is minimum in this case and hence chosen as the optimum value. Though it has maximum pixel error in the cover image, the difference from other β values is not as significant as seen in pixel errors of the secret image.

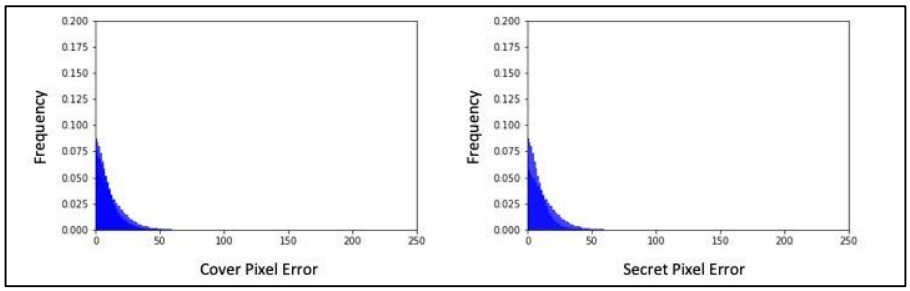


Figure 5: For β = 1, Left: Cover pixel error histogram, Right: Secret Pixel Error

The above graphs shown in Figure 5 represent the distribution of error in each pixel of both cover and secret image after encoding and decoding process for the most appropriate value of β i.e. 1.

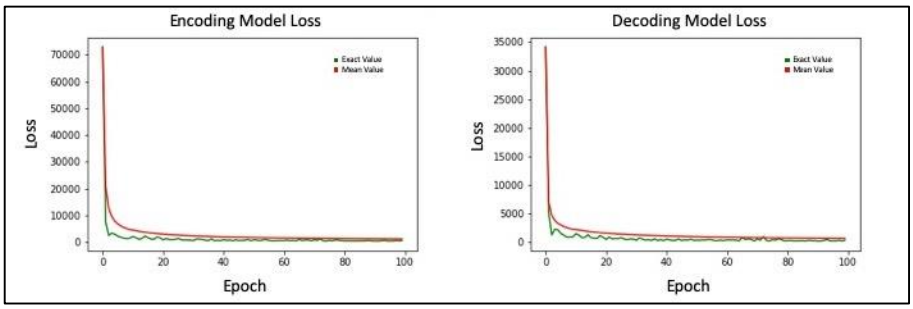


Figure 6: Left: Encryption Model Loss, Right: Decryption Model Loss

Figure 6 depicts the mean values for β = 1. The green and red on the line plot represent exact values and mean values respectively. Encoding Loss is the total loss of model that is the sum of squares of |C-C'| + β*|S-S'|. Reveal loss is the sum of squares of β*|S-S'|.

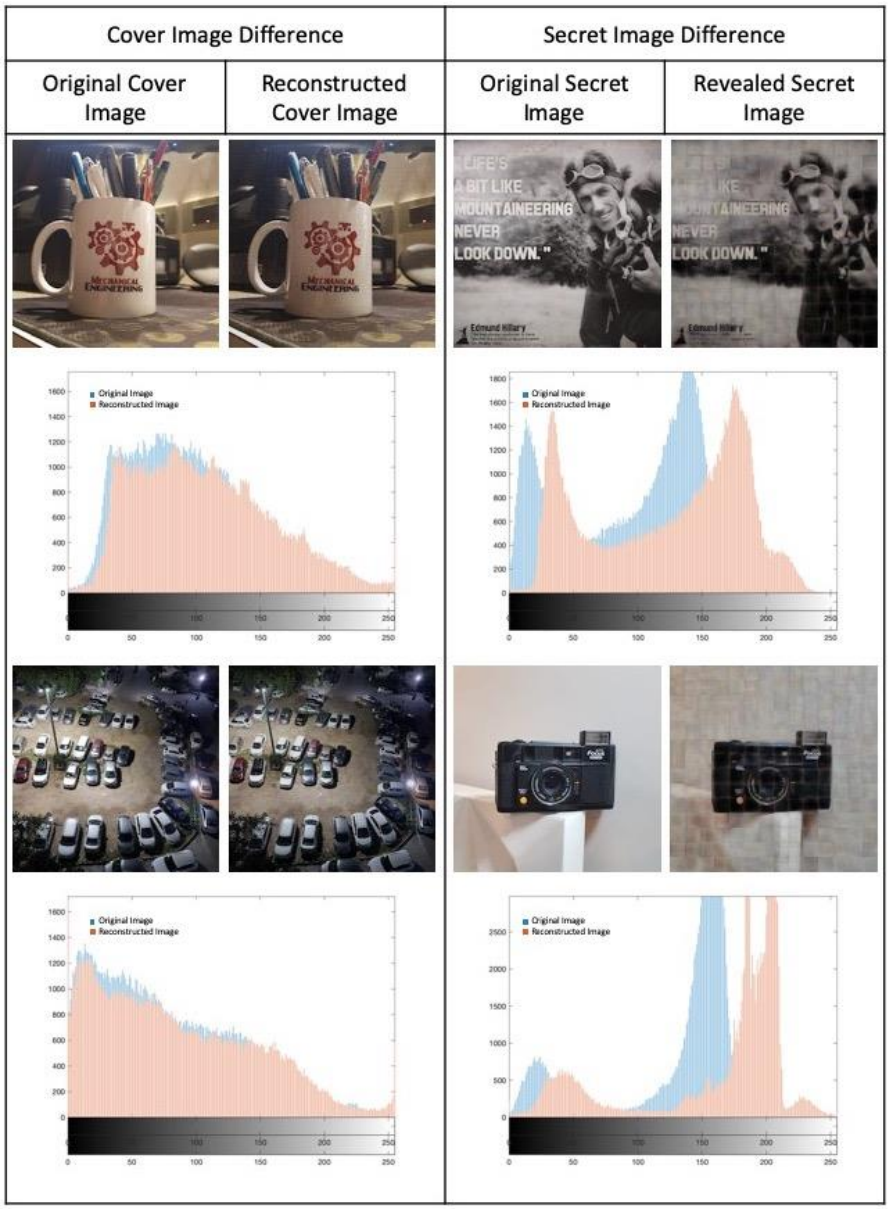


Figure 7: Representation of pixel densities in cover and secret image using histograms.

The histograms in Figure 7 represents the difference in pixel density in the original cover (left side) and secret (right side) images. Orange color represents the original cover and secret whereas blue color represents the reconstructed cover and revealed secret respectively. Due to adding an extra layer for cryptography, there is a tradeoff between secrecy and reveal loss.

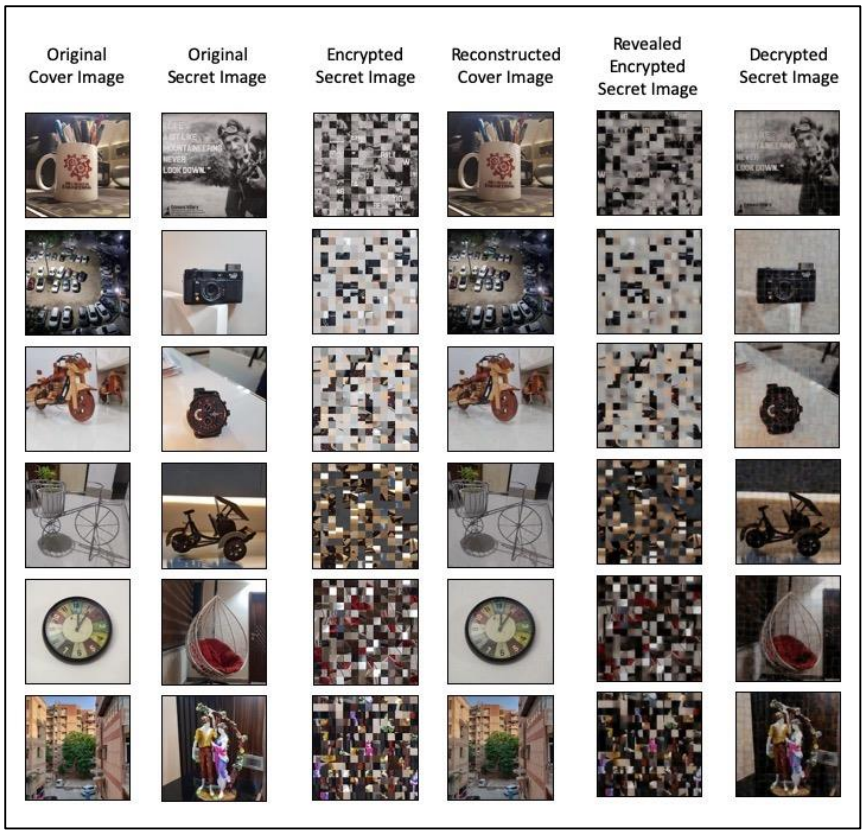


Figure 8: Results of integrating neural networks with an encryption layer. Left pair: original cover and secret image, Center pair: encrypted secret and reconstructed cover image, Right pair: revealed encrypted and decrypted the secret image.

As shown in Figure 8, the first column represents the original cover image used to hide the encrypted secret image shown in the third column. The fourth column represents the reconstructed cover image which has an embedded encrypted secret image and we finally get the decrypted secret image in the sixth column. The output represents that the secret image is successfully retrieved. Although, it contains some error the secret image is easily understandable. Figure 8 showcases only 6 tuples, but we were successfully able to hide and retrieve over 1000 images.

5. What If the Original Cover Image Is Available?

As explained by Baluja [11], the original cover image i.e. without the secret image embedded is almost impossible for someone to obtain. But what if it was obtained? After examining, without the decoding network secret could be partially be revealed. The secret image can be revealed by taking the difference between the original and reconstructed image i.e. (|S-S'|).

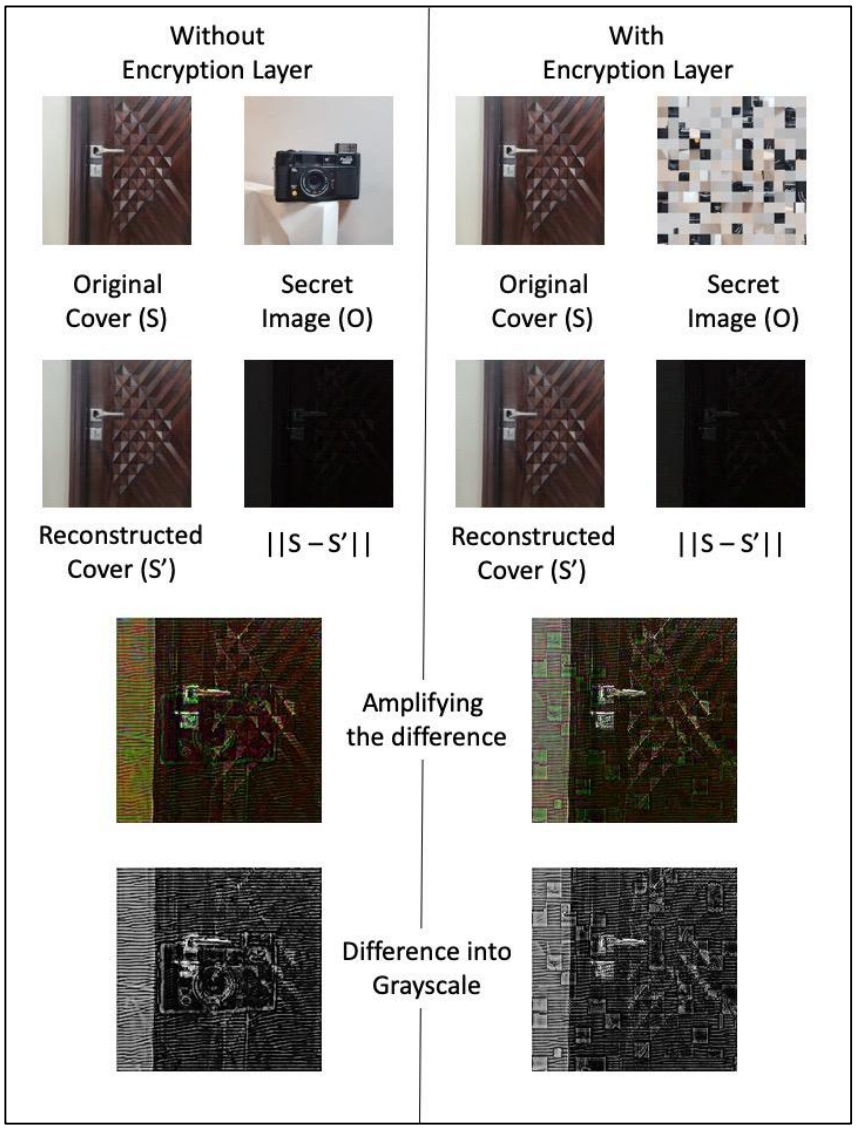


Figure 9: Effect of encryption layer.

As Figure 9 shows, without the first layer of encryption, data can be revealed by taking the difference between the original cover image and the reconstructed image. After obtaining the difference the secret is not visible but after enhancing the image and converting it into grayscale, the great resemblance between secret image and the residual image is found.

If we use the first layer of encryption, the secret is impossible to obtain as after enhancing and converting the residual image into grayscale we see only boxes and the residual image is not comprehensible which is depicted in Figure 9. Hence, it

makes the system robust and almost impossible for someone to extract the features of the secret image.

6. Conclusion

This paper propounds a new method for image steganography which is much more secure than the previous implementations. Although, it uses some common steganography techniques, integrating it with cryptography and neural networks make it arduous to break. The encryption layer added provides an additional layer of security with deep neural networks. As shown in Figure 8 we successfully embedded and revealed the secret image from the container. Further, the main reason to add an additional encryption layer was that if the original cover was made public, the secret will still remain secure. In earlier implementations, as shown in Figure 9 if the additional layer is not added, one can partially decrypt the sensitive information.

7. Future Work

In this section, we discuss possibilities for future work. Primarily, the method used for encryption can be replaced with some sophisticated encryption algorithms. Although the method for encrypting the secret image is secure, using cryptographic methods such as Advanced Encryption Standard (AES) or Data Encryption Standard (DES) is beyond the scope of this paper. The reason behind not using those algorithms is the problem faced when recovering the encrypted image due to lossy neural networks. In the future, if lossless neural networks are achieved, we would be able to implement them with modern cryptography techniques. The scope of building cryptographic algorithms that could work with neural networks is also a great field to explore in the future.

Moreover, integrating neural networks with image steganography creates a variety of possibilities in this domain. As stated by Gopalan (2002) [42] LSB steganography can be implemented with audio files as well. Therefore, we can similarly integrate neural nets with audio files. Not only LSB steganography there are numerous other steganography techniques with which neural networks can be integrated in the future.

Acknowledgement

We are very grateful to the editors and reviewers for their suggestions on our paper.

Conflicts of Interest

There is no conflict of interest.

References

[1] Provos, N., & Honeyman, P. (2003). Hide and seek: An introduction to steganography. IEEE security & privacy, 1(3), 32-44. https://doi.org/10.1109/MSECP.2003.1203220

[2] Johnson, N. F., & Jajodia, S. (1998). Exploring steganography: Seeing the unseen. Computer, 31(2), 26-34. https://doi.org/10.1109/MC.1998.4655281

[3] Morkel, Tayana & Eloff, Jan & Olivier, Martin. (2005). An overview of image steganography. 1-11.

[4] Fridrich, J., & Goljan, M. (2002, April). Practical steganalysis of digital images: state of the art. In Security and Watermarking of Multimedia Contents IV (Vol. 4675, pp. 1-13). International Society for Optics and Photonics. https://doi.org/10.1117/12.465263

[5] Shih, F. Y. (2017). Digital watermarking and steganography: fundamentals and techniques. CRC press. https://doi.org/10.1201/9781315219783

[6] Avcibas, I., Memon, N., & Sankur, B. (2003). Steganalysis using image quality metrics. IEEE transactions on Image Processing, 12(2), 221-229. https://doi.org/10.1109/TIP.2002.807363

[7] Li, B., He, J., Huang, J., & Shi, Y. Q. (2011). A survey on image steganography and steganalysis. Journal of Information Hiding and Multimedia Signal Processing, 2(2), 142-172.

[8] Subhedar, M. S., & Mankar, V. H. (2014). Current status and key issues in image steganography: A survey. Computer science review, 13, 95-113. https://doi.org/10.1016/j.cosrev.2014.09.001

[9] Fridrich, J., Goljan, M., & Du, R. (2001, October). Reliable detection of LSB steganography in color and grayscale images. In Proceedings of the 2001 workshop on Multimedia and security: new challenges (pp. 27-30). ACM. https://doi.org/10.1145/1232454.1232466

[10] Neeta, D., Snehal, K., & Jacobs, D. (2006, December). Implementation of LSB steganography and its evaluation for various bits. In 2006 1st International Conference on Digital Information Management (pp. 173-178). IEEE. https://doi.org/10.1109/ICDIM.2007.369349

[11] Baluja, Shumeet. "Hiding Images in Plain Sight: Deep Steganography." NIPS (2017).

[12] Anderson, R. J., & Petitcolas, F. A. (1998). On the limits of steganography. IEEE Journal on selected areas in communications, 16(4), 474-481. https://doi.org/10.1109/49.668971

[13] Marwaha, P., & Marwaha, P. (2010, July). Visual cryptographic steganography in images. In 2010 Second international conference on computing, communication and networking technologies (pp. 1-6). IEEE. https://doi.org/10.1109/ICCCNT.2010.5591730

[14] Djebbar, F., Ayad, B., Hamam, H., & Abed-Meraim, K. (2011, April). A view on latest audio steganography techniques. In 2011 International Conference on Innovations in Information Technology (pp. 409-414). IEEE. https://doi.org/10.1109/INNOVATIONS.2011.5893859

[15] Douglas, M., Bailey, K., Leeney, M., & Curran, K. (2018). An overview of

steganography techniques applied to the protection of biometric data. Multimedia Tools and Applications, 77(13), 17333-17373. https://doi.org/10.1007/s11042-017-5308-3

[16] Jiang, N., Zhao, N., & Wang, L. (2016). LSB based quantum image steganography algorithm. International Journal of Theoretical Physics, 55(1), 107-123. https://doi.org/10.1007/s10773-015-2640-0

[17] Wang, H., & Wang, S. (2004). Cyber warfare: steganography vs. steganalysis. Communications of the ACM, 47(10), 76-82. https://doi.org/10.1145/1022594.1022597

[18] Basheer, I. A., & Hajmeer, M. (2000). Artificial neural networks: fundamentals, computing, design, and application. Journal of microbiological methods, 43(1), 3-31. https://doi.org/10.1016/S0167-7012(00)00201-3

[19] Yao, X. (1999). Evolving artificial neural networks. Proceedings of the IEEE, 87(9), 1423-1447. https://doi.org/10.1109/5.784219

[20] Zurada, J. M. (1992). Introduction to artificial neural systems (Vol. 8). St. Paul: West publishing company.

[21] Gardner, M. W., & Dorling, S. R. (1998). Artificial neural networks (the multilayer perceptron)-a review of applications in the atmospheric sciences. Atmospheric environment, 32(14-15), 2627-2636. https://doi.org/10.1016/S1352-2310(97)00447-0

[22] O'Shea, Keiron & Nash, Ryan. (2015). An Introduction to Convolutional Neural Networks. ArXiv e-prints.

[23] Tiwari, Prayag & Qian, Jia & Li, Qiuchi & Wang, Benyou & Gupta, Deepak & Khanna, Ashish & Rodrigues, Joel & Albuquerque, V.H.C.. (2018). Detection of Subtype Blood Cells using Deep Learning. Cognitive Systems Research. 52. 10.1016/j.cogsys.2018.08.022. https://doi.org/10.1016/j.cogsys.2018.08.022

[24] Khamparia, Aditya & Singh, Aman & Anand, Divya & Gupta, Deepak & Khanna, Ashish & Kumar, N Arun & Tan, Joseph. (2018). A novel deep learning-based multi-model ensemble method for the prediction of neuromuscular disorders. Neural Computing and Applications. 10.1007/s00521-018-3896-0. https://doi.org/10.1007/s00521-018-3896-0

[25] Khamparia, Aditya & Gupta, Deepak & Nhu, Nguyen & Khanna, Ashish & Pandey, Babita & Tiwari, Prayag. (2019). Sound Classification Using Convolutional Neural Network and Tensor Deep Stacking Network. IEEE Access. PP. 1-1. 10.1109/ACCESS.2018.2888882. https://doi.org/10.1109/ACCESS.2018.2888882

[26] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

[27] Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882. https://doi.org/10.3115/v1/D14-1181

[28] Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2414-2423). https://doi.org/10.1109/CVPR.2016.265

[29] Jaderberg, M., Vedaldi, A., & Zisserman, A. (2014). Speeding up convolutional

neural networks with low rank expansions. arXiv preprint arXiv:1405.3866. https://doi.org/10.5244/C.28.88

[30] LeCun, Y., Kavukcuoglu, K., & Farabet, C. (2010, May). Convolutional networks and applications in vision. In Proceedings of 2010 IEEE International Symposium on Circuits and Systems (pp. 253-256). IEEE. https://doi.org/10.1109/ISCAS.2010.5537907

[31] Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. science, 313(5786), 504-507. https://doi.org/10.1126/science.1127647

[32] Kingma, Diederik & Ba, Jimmy. (2014). Adam: A Method for Stochastic Optimization. International Conference on Learning Representations.

[33] Kumar, Mohit & Agrawal, Akshat & Garg, Ankit. (2014). A Review on Various Digital Image Encryption Techniques and Security Criteria. International Journal of Computer Applications. 13. 10.5120/16854-6720. https://doi.org/10.5120/16854-6720

[34] Kaur, Rajinder. (2013). Image Encryption Techniques: A Selected Review. IOSR Journal of Computer Engineering. 9. 80-83. 10.9790/0661-0968083. https://doi.org/10.9790/0661-0968083

[35] Srivastava, Abhinav. "A survey report on Different Techniques of Image Encryption." (2012).

[36] Li, C. G., Han, Z. Z., & Zhang, H. R. (2002). Image encryption techniques: A survey [j]. Journal of Computer Research and Development, 10(023), 15.

[37] Khan, M., & Shah, T. (2014). A literature review on image encryption techniques. 3D Research, 5(4), 29. https://doi.org/10.1007/s13319-014-0029-0

[38] Geetha, P. & Vs, Jayanthi & A.N., Jayanthi. (2018). Optimal Visual Cryptographic Scheme with multiple share creation for Multimedia Applications. Computers & Security. 78. 10.1016/j.cose.2018.07.009. https://doi.org/10.1016/j.cose.2018.07.009

[39] Ye, G. (2010). Image scrambling encryption algorithm of pixel bit based on chaos map. Pattern Recognition Letters, 31(5), 347-354. https://doi.org/10.1016/j.patrec.2009.11.008

[40] Amnesh Goel, Nidhi Chandra "A Technique for Image Encryption Based On Explosive n*n Block displacement Followed By Inter-Pixel Displacement of RGB Attribute of A Pixel" 2012 International Conference on Communication Systems and Network Technologies, 11-13 May 2012, page: 884-888. https://doi.org/10.1109/CSNT.2012.190

[41] Peter Young, Alice Lai, Micah Hodosh and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions, Transactions of the Association for Computational Linguistics, 2(Feb):67-78, 2014. https://doi.org/10.1162/tacl_a_00166

[42] Gopalan, K. (2003, July). Audio steganography using bit modification. In 2003 International Conference on Multimedia and Expo. ICME'03. Proceedings (Cat. No. 03TH8698) (Vol. 1, pp. I-629). IEEE. https://doi.org/10.1109/ICME.2003.1220996

Wireless Personal Communications
https://doi.org/10.1007/s11277-019-07018-1

High Diversity Gain MIMO-Antenna for UWB Application with WLAN Notch Band Characteristic Including Human Interface Devices

Gaurav Saxena[1,2] · Priyanka Jain[1] · Yogendra Kumar Awasthi[3]

© Springer Science+Business Media, LLC, part of Springer Nature 2019

Abstract
In this paper, a UWB–MIMO antenna with the WLAN band-notch (5.1–5.85 GHz) characteristic is offered. This antenna consists of two radiated patch feeding with a tapered line and fabricated on abundantly available FR-4 substrate having the size of $36\times22\times1.6$ mm^3. A notched-band response is achieved by introducing an open-ended stub on the ground plane. Results illustrate that the designed antenna has an impedance bandwidth from 3.1 to 11.2 GHz as well as good isolation i.e. $S_{21}\leq-30$ dB. Radiation efficiency is greater than 0.75 except the notched band is <0.5. Diversity performance is also set the new paradigm in terms of ECC (≤0.008), TARC (≤-25 dB), CCL (≤0.3 bits/s/Hz), Mean effective gain ratio (MEGi) $\cong1$ and directive gain (≥9.95 dB) except the notched band. Proposed antenna characteristics are also found suitable for a human interface device, low-cost, and easily fabricated. Simulated results of the proposed antenna are tested and verified by the experimental results.

Keywords Ultra-wideband (UWB) · Multiple input multiple output (MIMO) · Wireless local area network (WLAN) · Diversity gain (DG) · Envelope correlation coefficient (ECC) · Channel capacity loss (CCL) · Total active reflection co-efficient (TARC)

✉ Priyanka Jain
priyankajain@dtu.ac.in

Gaurav Saxena
gaurav.saxena@galgotiacollege.edu

Yogendra Kumar Awasthi
yogendra@mru.edu.in

[1] Department of Electronics and Communication Engineering, Delhi Technological University, Delhi 110042, India

[2] Galgotia College of Engineering and Technology, Greater Noida 201306, India

[3] Antenna Fabrication and Measurement Laboratory, Electronics and Communication Engineering, Manav Rachna International Institute of Research and Studies, Faridabad 121004, India

Published online: 24 December 2019

1 Introduction

MIMO antenna is extensively utilized in contemporary wireless RF communication system to improve the diversity problem in a single input single output-SISO antenna. The UWB system operates in a wide range of frequency among small spectral power. The signal to noise ratio is also less at the receiver end. That's why; band-notch is the only opportunity to mitigate the interference from UWB applications through wireless protocols like WiMAX, LTE, and WLAN. The electromagnetic bandgap is used to improve the isolation of the MIMO antenna and co-design approach provides better diversity results [1, 2]. Isolation of MIMO is improved by introducing carbon black film, which can be absorbing the intrusion among antenna elements [3]. Moreover, high isolation is attained by T-shaped slot and stub on the ground plane and dual-polarization (Linear to circular) is achieved by using two perpendicular feeding lines at the patch [4, 5]. In addition, to achieve good isolation performance, T-shaped protuberant ground and Y-shaped quadrate patch are suggested for UWB applications [6]. An extruded T-shaped decoupling structure is introduced in UWB–MIMO antenna to achieve a good diversity performance [7] and orthogonal monopole radiators are also used to attain high isolation without introducing decoupling structure in the ground [8]. Furthermore, by using a T-shaped slot and two-line slots on the ground, high isolation and impedance bandwidth have been achieved [3]. Moreover, inverted L-shaped slits are introduced to produce good diversity performance [9].

By using Monopole antennas, it is achieved UWB operation from 3.1 to 10.6 GHz. In addition, T-shaped slot and stub in the ground are used to minimize the mutual coupling by − 15 dB and two ground-strips are produced an excellent band-notch characteristic for WLAN (5.15–5.85 GHz) to suppress interference [10]. A tapered feed microstrip line with two identical triangular shaped patch radiators and two triangles with J-shaped slit ground are built to get the WLAN notch-band characteristic from 5.10 to 5.85 GHz and INSAT and extended C-band from 6.7 to 7.1 GHz [11]. Henceforth, folded printed monopole antenna is designed to enhance the bandwidth, which is joined with an inverted-L parasitic element and introduced an open-stub to eliminate the WLAN (5.15–5.85 GHz) band [12, 13]. A rectangle shaped stub and split ring resonator (SRR) are providing to attain good isolation and rejection of band simultaneously. Results have shown improved matching of impedance with the required radiation pattern and constant gain also achieved in the operating bandwidth [14]. Moreover, a compact rectangular dielectric resonator (DRA) MIMO antenna with microstrip feed is designed to achieve good bandwidth and inserted two stubs in the ground to get high isolation between antennas [15]. The wireless communication system has various generations (1G–5G), which is supported by SISO to MIMO systems and large bandwidth of the system is one of the foremost factors for high transmission data rate, to enhance the services like triple play i.e. data, audio, and video. Hence, at user end the capacity directly influence the communication system quality. Although MIMO system provides these features in a challenging world to provide high data rate, low channel capacity loss and high signal to noise ratio (SNR) [16] and orthogonal radiating staircase shaped patches with modified co-planer waveguide feed MIMO antenna is designed with 5.5 GHz band-notched characteristics [17]. To provide high isolation and UWB bandwidth with notched band applications, MIMO antennas are presented [18–20]. Henceforth, two antennas placed orthogonally to each other and decoupling structure designed in the ground plane to accomplish better isolation between two radiator elements. To improve isolation in MIMO antenna system, two defective complementary open-loop resonators also used to get required results between antennas [21, 22].

In this paper, two-element tapered fed high diversity gain, WLAN notched UWB–MIMO antenna is proposed and it has fabricated on FR-4 substrate with a compact size of $36\times22\times1.6$ mm^3. A band-notch characteristic is obtained through two open-ended stubs on the ground. Furthermore, the diversity characteristic of "mean effective gain" MEGi/MEGj ratio and channel capacity loss (CCL) has also studied. Section 2 presents the proposed antenna design procedure. Section 3 elucidates the proposed antenna results and discussion and Section 4 depicts the diversity performance of the proposed antenna. The achieved results of measurement and simulation are in better concurrence.

2 MIMO Antenna Design Procedure

Numerous wideband applications such as microwave medical imaging, ground penetrating radar (GPR) and RFID tag for inventory control and asset management require a compact and inexpensive UWB–MIMO wireless transceivers. UWB–MIMO base station receiver architecture with different blocks is shown in Fig. 1.

Impulse data is transmitted using UWB–MIMO antenna and after that UWB–MIMO receiver antenna receives the data in a different path and produces the base-band signal, where the information is first encoded, digitally modulated, and then converted into pulse using pulse generator. This UWB–MIMO antenna has shown in Fig. 2a and b with WLAN notch is fabricated on FR-4 substrate of height (h) 1.6 mm with relative permittivity (ε_r) of 4.4, loss tangent (tanδ) of 0.02 by the dimension of 36 mm×22 mm. The antenna is designed by a circular with triangular notched patches and to enhance the isolation between two radiated patches, a T–I shaped slot in the ground is groomed, which improves the isolation better than −30 dB in wished-for UWB bandwidth. The resonant frequency of circular patch for TM_{11} mode is 7.15 GHz which is used to calculate the patch radius to achieve desired bandwidth (3.1–11.2 GHz).

UWB bandwidth in MIMO antenna including notch band characteristic is achieved by using a tapered feed line together with the partial tailored ground plane. The fundamental resonance frequency of the monopole antenna is prearranged by Eq. (1).

$$f_r = \frac{144}{l_1 + d_1 + h + \frac{A_1}{2\pi\sqrt{\epsilon_r+1}} + \frac{A_2}{2\pi\sqrt{\epsilon_r+1}}} \tag{1}$$

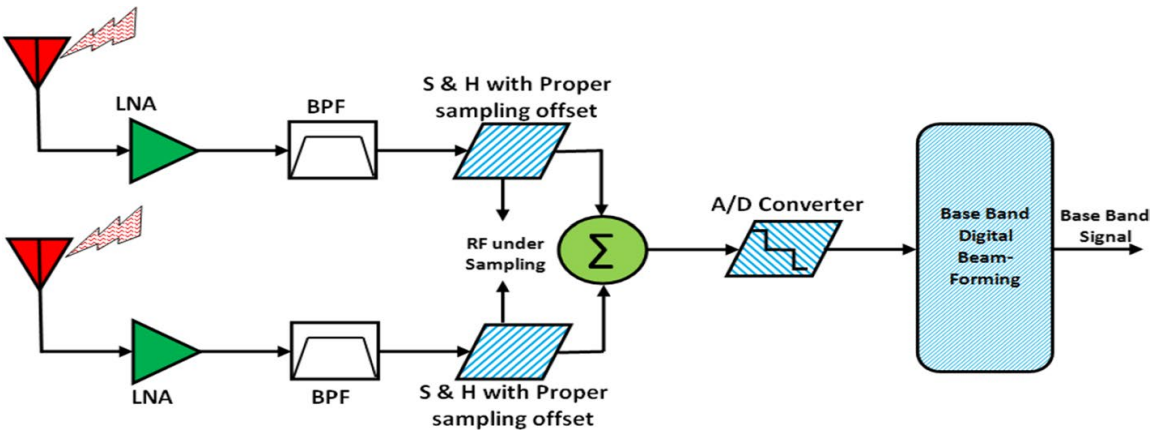


Fig. 1 A generic block diagram of two-element UWB–MIMO antenna-based receiver station

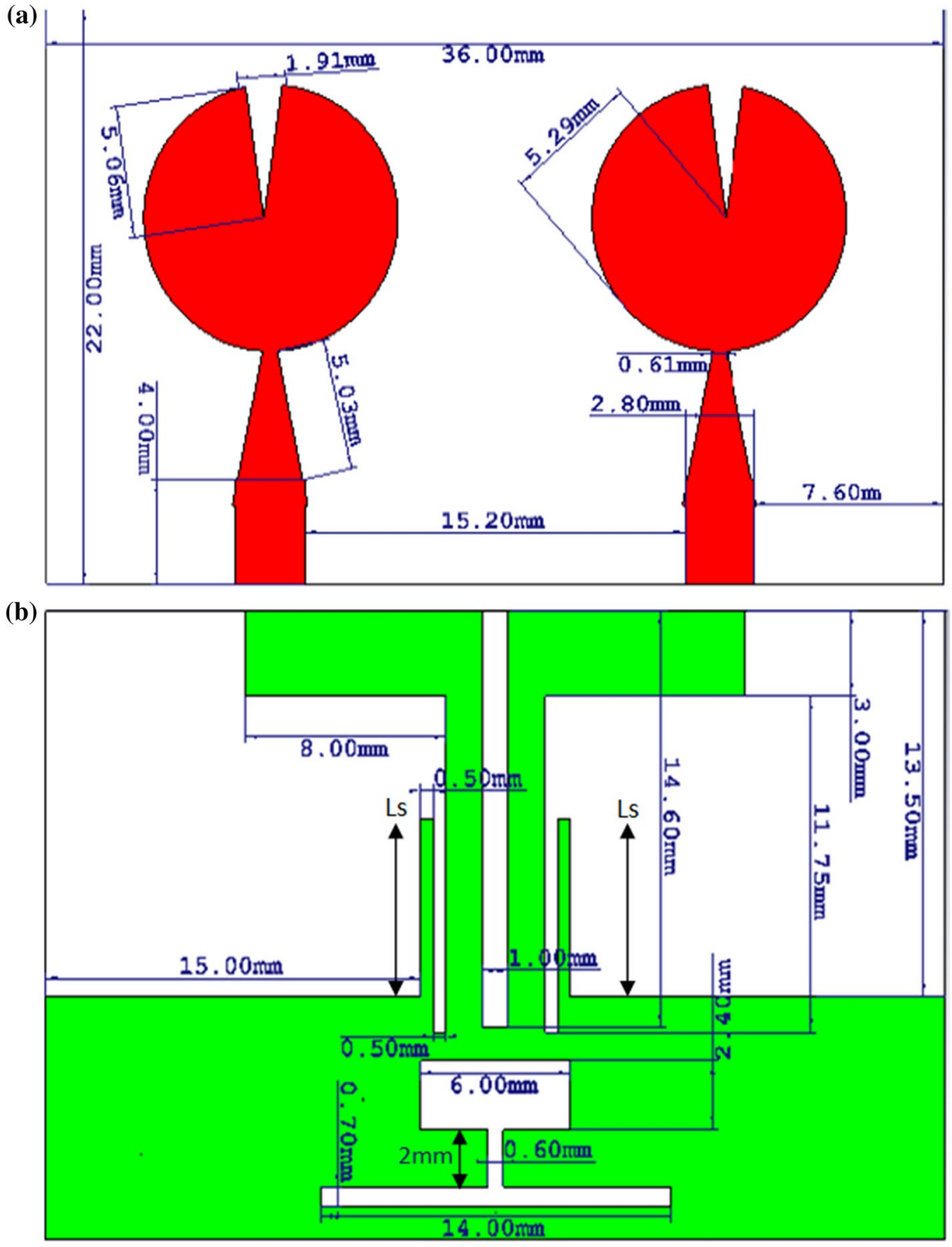


Fig. 2 Two-element UWB planar monopole MIMO antenna **a** front view, **b** bottom view

where A_1, A_2 is the area of anticipated radiation patches and ground plane, l_1, length of the ground plane, d_1, diameter of circular radiated patches. Step by step procedure of the proposed UWB-MIMO antenna design is ushered in Fig. 3 and corresponding return loss and isolation are presented in Fig. 4a and b respectively. Antenna-1 provides unsatisfactory return loss and isolation in intentional UWB range. By changing the feeding line width of antenna-1, return loss and isolation is improved as depicted in antenna-2.

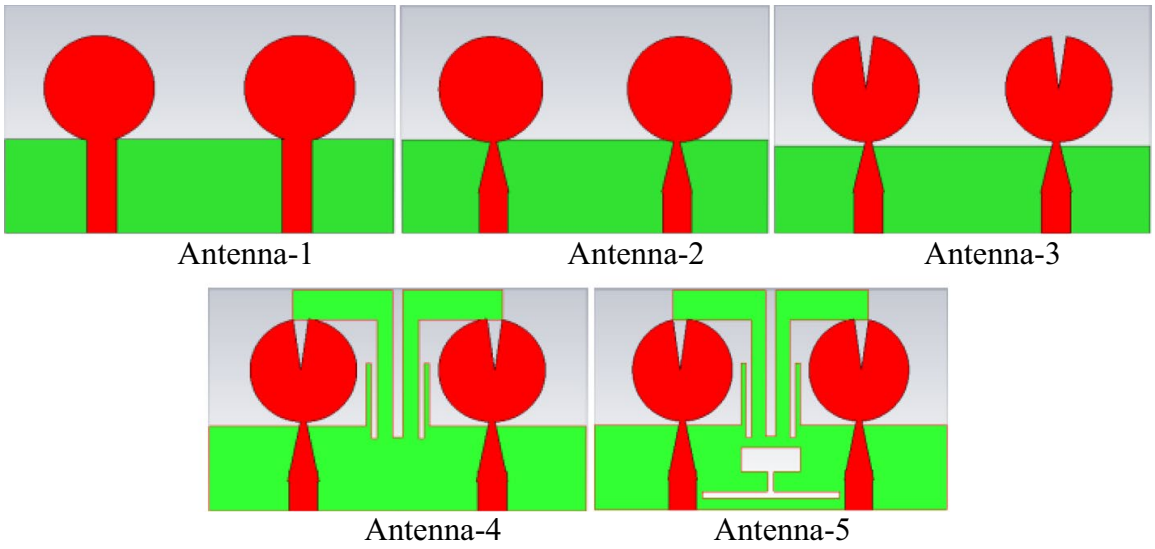


Fig. 3 Geometrical design steps of a designed antenna with WLAN band-notch characteristic

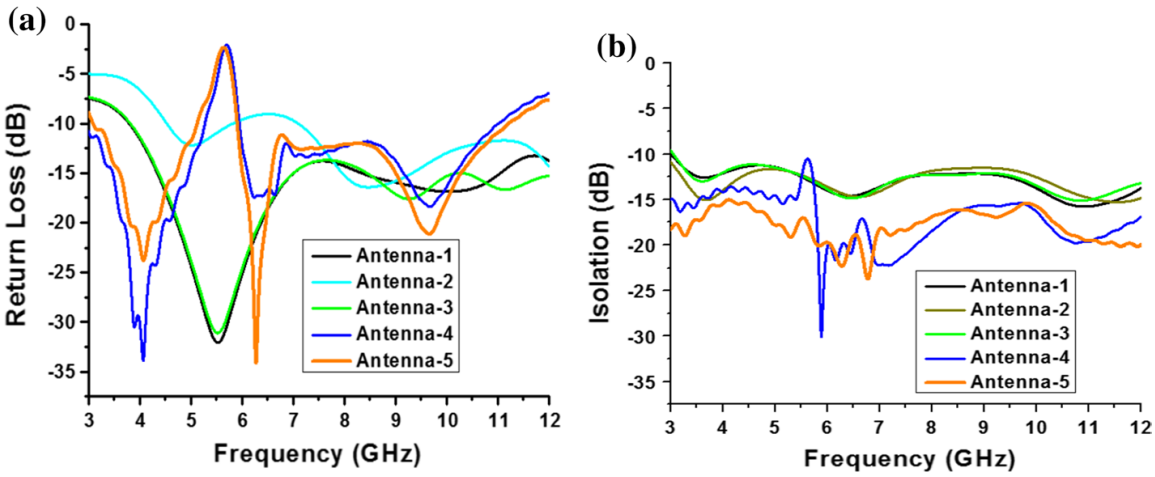


Fig. 4 Simulated (using CST MWS) results of **a** return loss, **b** isolation of the antennas shown in Fig. 3

In antenna-3, a triangular shape etched from patches which improve isolation as well as return loss. Hence, an antenna-4, impart vertical stubs on the ground plane to get better return loss from 3.1 to 4.8 GHz and it also provides a notch band characteristics for WLAN (5.1–5.9 GHz) although poor isolation in the intended lower frequency region. Therefore, to improve isolation in entire UWB bandwidth, fixed another T and I shaped slot in ground, the same as shown in antenna-5.

The notched band can be achieved by varying the ground stubs length (L_s) at the notch frequency, $f_n = c/\left(2 \times L_s \times \sqrt{\varepsilon_{reff}}\right)$, where c is the speed of the signal in a vacuum (3×10^8 m/s), ε_{reff} is the effective relative permittivity [23] of the substrate and approximated length of L_s is 6.3 mm at notched frequency.

Frequency selectivity of the rejection band is an important parameter to design a UWB antenna with notch band. For selectivity of the notched band of the UWB–MIMO antenna, another ubiquitous parameter is 'Roll off rate' i.e. control required a bandwidth of the WLAN band. Furthermore, RoC-Roll off criteria is the ratio of the bandwidth at − 5 dB and − 10 dB, i.e. $RoC = BW_{-5\ dB}/BW_{-10\ dB}$ and the RoC of proposed antenna is 0.44, which is acceptable for practical design. The bandwidth of notched

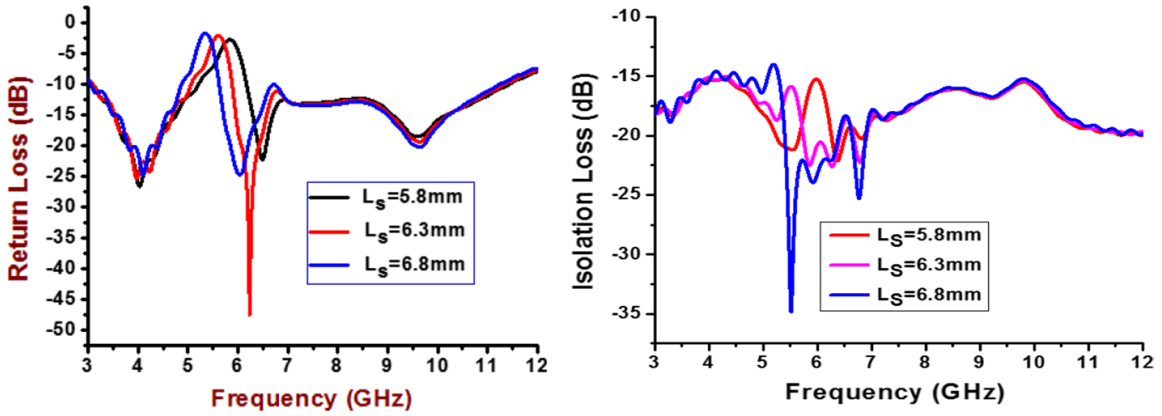


Fig. 5 Simulated results of return loss and Isolation of antenna with a variation of length of stubs (L_s) in the ground plane

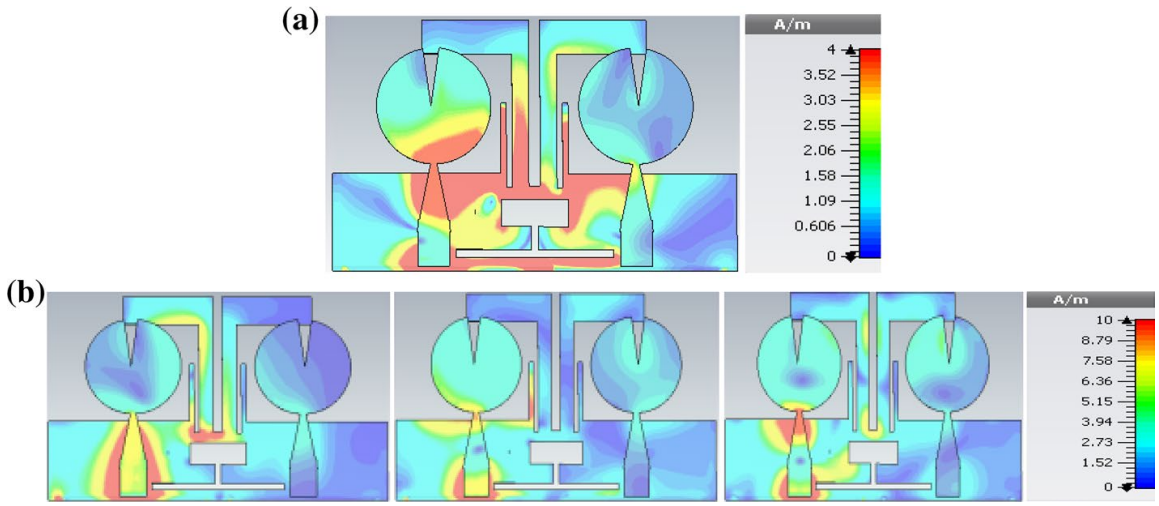


Fig. 6 Surface current distributions at **a** notch band frequency i.e. 5.6 GHz and **b** 4.1, 6.6 and 9.6 GHz when port-2 terminated with 50 Ω impedance and port-1 is excited

band is also controlled by the length of stubs (Ls) in ground plane and the results of return loss, centered at 5.6 GHz as made known in Fig. 5. By varying Ls from 5.8 to 6.8 mm, return loss curve shifts from higher to lower frequency and value of return loss at the intended notch (6.3 mm) is − 2.0 dB. Impedance bandwidth is also maintained significantly, in the entire bandwidth throughout optimization of L_s.

It is hardnosed from Fig. 6a, where strong current accumulates around vertical stubs in the ground plane, which produces a notched for WLAN band in UWB bandwidth to mitigate interference. Figure 6b shows the distribution of current on the antenna at various frequencies like 4.1, 6.6 and 9.6 GHz respectively. It is too observed that the flows of current on the tapered feed line to circular element while a small current pours across the second element of MIMO. The slot in ground plane works as an isolator between the two elements and this is preventing the coupling of the electromagnetic energy.

3 Results and Discussion

The scattering parameter results (measured and simulated) are shown in Fig. 7. Eventually, measured S_{12} shows better isolation than the simulation results and isolation between two antenna radiators is <-30 dB for the entire UWB bandwidth. Results illustrate that the return loss of proposed MIMO antenna is ≤ -10 dB except for notched band (WLAN) and bandwidth of the antenna is 8.1 GHz (from 3.1 to 11.2 GHz), that fulfills the FCC requirement for UWB applications without interference.

Maximum peak gain of antenna has revealed the better conformity between simulated and measured results and the maximum value of gain is 4.85 dB. The radiation efficiency of antenna throughout the requisite band is higher than the permissible limit except for the notched band, is centered at 5.6 GHz. Consequently, proposed MIMO can appropriate to work in the high interference surrounding as presented in Fig. 7b.

Figure 8a–c shows the 3D radiation pattern of proposed MIMO antenna at 4.1, 6.6 and 9.6 GHz frequency correspondingly, when port-1 is excited and port-2 is terminated with 50 Ω or vice versa. Radiation patterns are slightly directional because power radiated towards the left of the yz-plane when port-1 is excited and when port-2 of MIMO antenna is excited, the power is radiating towards the right of the yz-plane to achieved pattern diversity performance of an antenna. Co-polarization is the power level of an electromagnetic wave in an intended direction on the other hand cross-polarization is the power level of an electromagnetic wave in an un-intended direction due to interference at reception or transmission antenna. Co-polarization and Cross-polarization at different frequencies are calculated by the simulator (CST-Microwave studio) and validate these results by Vector Network Analyzer so we use two terms simulated, by CST-Microwave studio and measured by VNA. The prototype of proposed MIMO antenna on FR-4 substrate is fabricated by conventional printed circuit board design procedure and antenna under test in an anechoic chamber associated with VNA for measurement of radiation characteristics as shown in Fig. 8d and e.

Figure 9 shows the simulated/measured yz-plane (left) and xz-plane (right) radiation pattern of antenna at three different frequencies as 4.1, 6.5 and 9.6 GHz respectively with an acceptable agreement. Whereas the UWB–MIMO antenna consists of two indistinguishable radiated antenna elements and so the time of measurement of the 2D radiation pattern, port-1 is excited and port-2 is terminated by the 50 Ω load. It is observed that at operating

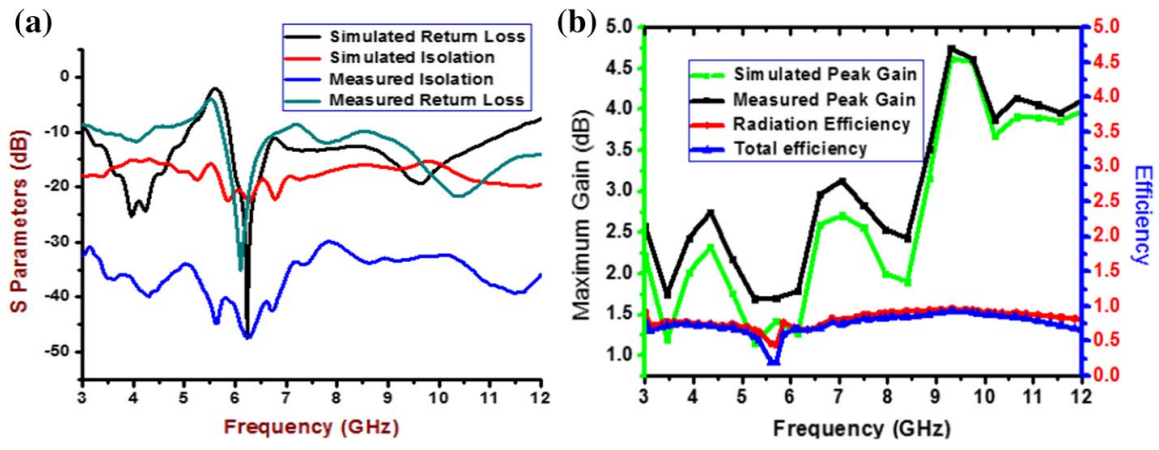


Fig. 7 Simulated and measured **a** S-parameter (Return and Isolation) **b** peak gain, radiation efficiency and total radiation efficiency of UWB–MIMO antenna

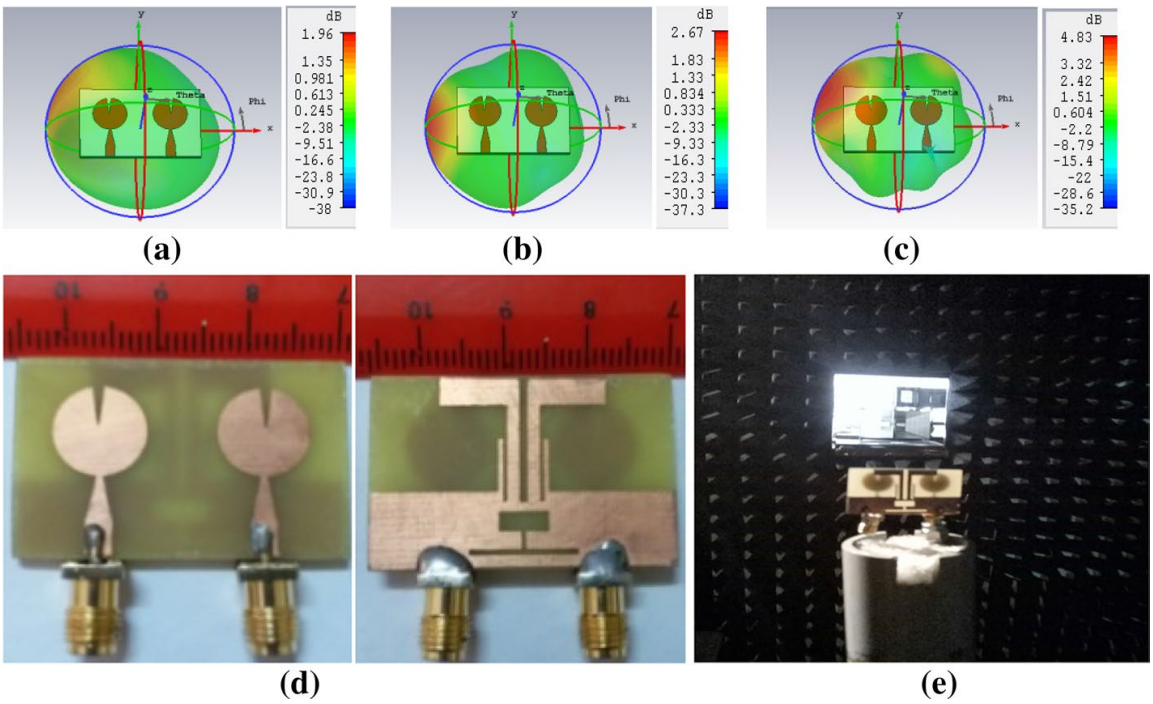


Fig. 8 The 3D radiation pattern of MIMO antenna at **a** 4.1 GHz, **b** 6.6 GHz, **c** 9.6 GHz only port-1 is excited and port-2 terminated by 50 Ω load, **d** fabricated prototype, **e** antenna under test in an anechoic chamber for radiation pattern measurement

frequency band the radiation patterns are omnidirectional in H (xz)-plane and dipole like in E (yz)-plane except for WLAN band. The splitting of the radiation lobes is also observed due to higher-order mode propagation at higher frequencies.

Specific absorption rate: SAR is a vital parameter in the antenna at which the human body is absorbed electromagnetic field energy when exposed, means the power absorbed per mass of body/head tissue as shown Fig. 10. It is typically averaged either over a small sample volume (typically 1 g or 10 g of tissue) or over the entire body. The SAR value of any RF and microwave electronic devices is obtained by penetration of electric field intensity in the human head tissue at near field environment [24, 25]. Furthermore, SAR estimation on the head at 4.1, 6.6 and 9.6 GHz frequency, the skin radius, bone radius, and brain radius are 79.324 mm, 75 mm and 65 mm respectively are used and other parameters are also listed in Table 1. The calculated value of SAR is less than 1.6 W/kg for the proposed antenna design as shown in Table 2 but this SAR performance investigates without any plastic jacket or cover, thus when we wrap proposed antenna into a plastic jacket or cover then SAR performance may be increased more. SAR for electromagnet energy is calculated by Eq. (2).

$$\mathrm{SAR} = \frac{1}{\mathrm{V}} \int \frac{\sigma(\mathrm{r})|\mathrm{E}(\mathrm{r})|^2}{\rho(\mathrm{r})} \mathrm{dr} \tag{2}$$

where σ = thermal conductivity of sample in S/m, E_{rms} = electric field in V/m, ρ = sample density in kg/m^3, V = volume of the sample in m^3.

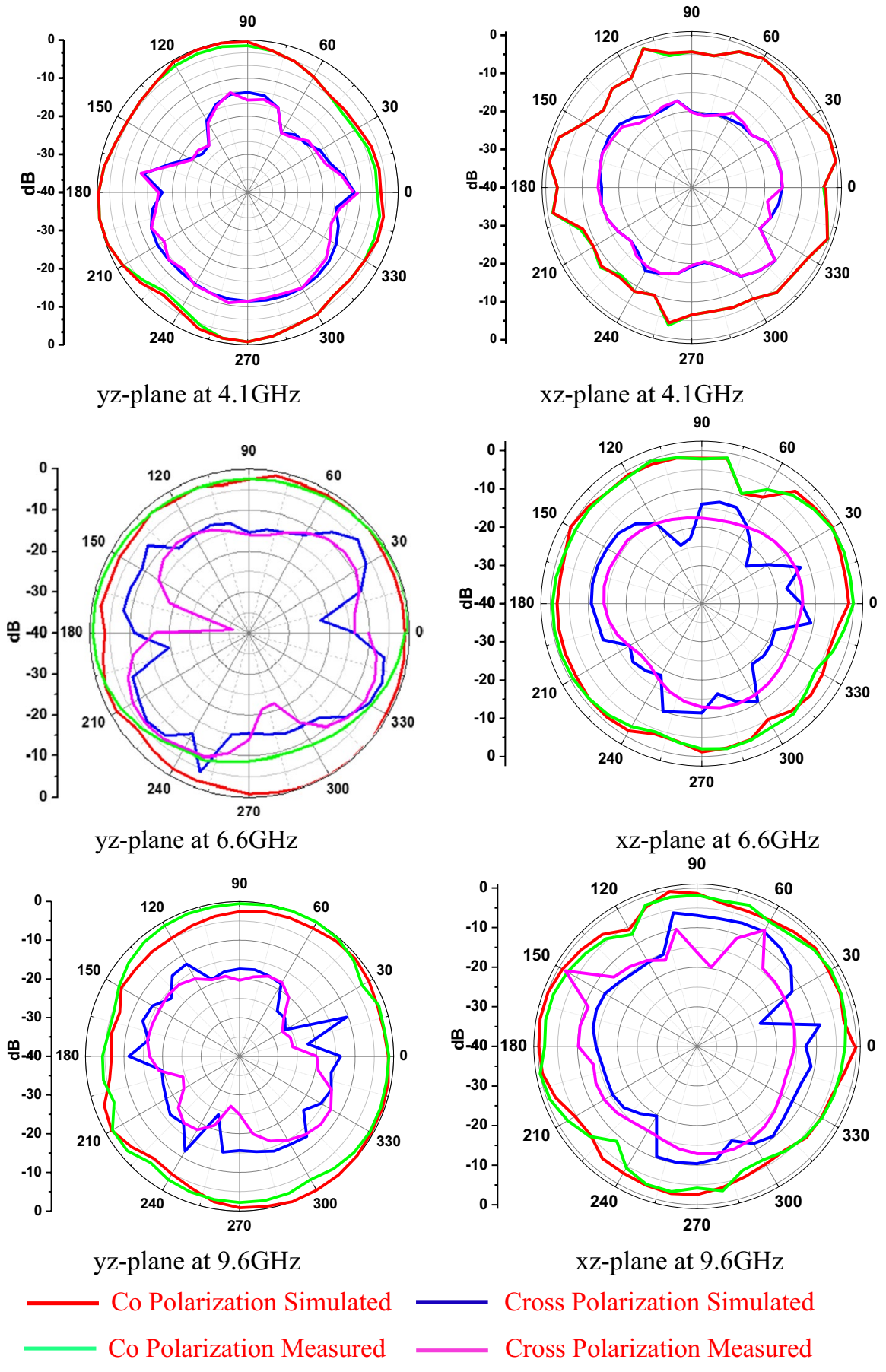


Fig. 9 Measured and simulated 2D radiation pattern at 4.1, 6.6 and 9.6 GHz for yz-plane (left) and xz-plane (right)

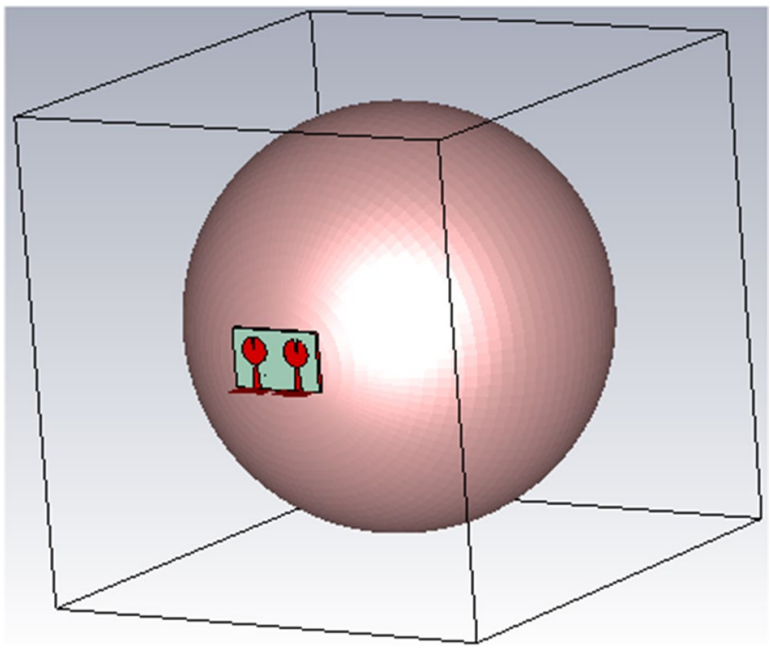


Fig. 10 Calculation of specific absorption rate near the human head in presence of the proposed MIMO antenna

Table 1 Parameters of the human head for calculation of specific absorption rate (SAR) by the proposed antenna at constant permeability (i.e. 1)

Type	Density (ρ) (kg/m^3)	Thermal conductivity (σ) (S/m)	Heat capacity (kJ/K/kg)	Blood flow (W/K/m^3)	Metabolism rate (W/m^3)
Bio-tissue skin	1100	0.293	3.5	9100	1620
Bio-tissue bone	1850	0.41	1.3	3400	610
Bio-tissue brain	1030	1.13	3.675	40,000	7100

Table 2 SAR values near the human head at different resonant frequencies when the distance between the human head and MIMO antenna is 13 mm

Specific absorption rate (W/kg)					
4.1 GHz		6.6 GHz		9.6 GHz	
1 (g)	10 (g)	1 (g)	10 (g)	1 (g)	10 (g)
1.28	0.86	1.39	0.93	1.49	1.04

4 MIMO Diversity Performance

MIMO diversity performance of the proposed antenna is evaluated in terms of Envelope Correlation Coefficient, Total Active Reflection Co-efficient, Diversity Gain, Multiplexing Efficiency and Channel Capacity Loss. The mutual coupling between two radiation patches and the amount of correlation between them is examined by ECC. It is calculated through using S-parameters by following Eq. (3a) and this equation is valid for measuring of ECC when the antenna is lossless and radiated power is uniformly distributed along with the antennas.

$$\mathrm{ECC} = \frac{\left|S_{11}^{*}S_{12} + S_{21}^{*}S_{22}\right|^{2}}{\left(1 - \left|S_{11}\right|^{2} - \left|S_{21}\right|^{2}\right) \times \left(1 - \left|S_{22}\right|^{2} - \left|S_{12}\right|^{2}\right)} \tag{3a}$$

But lossless is a hypothetical result which varies from empirical result so, the computation the ECC of the MIMO antenna by Eq. (3b) in terms of radiated fields

$$\mathrm{ECC} = \frac{\left|\iint_{4\pi} \left[E_i(\theta,\phi) \times E_j(\theta,\phi)\right] d\Omega\right|^{2}}{\iint_{4\pi} \left|E_i(\theta,\phi)\right|^{2} d\Omega \times \iint_{4\pi} \left|E_j(\theta,\phi)\right|^{2} d\Omega} \tag{3b}$$

where E_i and E_j are the radiated electric fields vector of the ith and jth elements of the MIMO antenna system and Ω denotes beam area.

The ideal value of ECC is zero for uncorrelated MIMO antenna but their acceptable value for practical MIMO antenna is ≤0.5 and as per Fig. 11a the value of ECC of designed MIMO antenna is less than 0.01 in UWB bandwidth excluding notched band. Hence, the above result specifies pretty low correlation between the two antenna ports, which shows the excellent diversity performance. The diversity gain can be calculated by $\mathrm{DG} = 10 \times \sqrt{\left(1 - \mathrm{ECC}^{2}\right)}$ and for satisfactory operation of MIMO antenna diversity gain should be closed to 10 dB and the value of diversity gain of UWB-MIMO antenna is 9.95 dB apart from the notched band as depicted in Fig. 11a.

In addition, to optimize the antenna channel capacity, the multiplexing efficiency is a maneuver parameter which is not only explored the total antenna efficiency, but also the co-relation and efficiency disparity. In addition, multiplexing efficiency (η_{MUX}) is the ratio of the power requirement of an antenna to the reference antenna. Figure 11b shows the multiplexing efficiency of the MIMO antenna and η_{MUX} is calculated by $\eta_{MUX} = \sqrt{\left(1 - \left|\rho_c\right|^{2}\right)\eta_1\eta_2}$, where ρ_c is a complex correlation between two side by side antenna i.e. ECC = ρ_c^2 and η_1, η_2 is the antenna total efficiency of MIMO antenna elements.

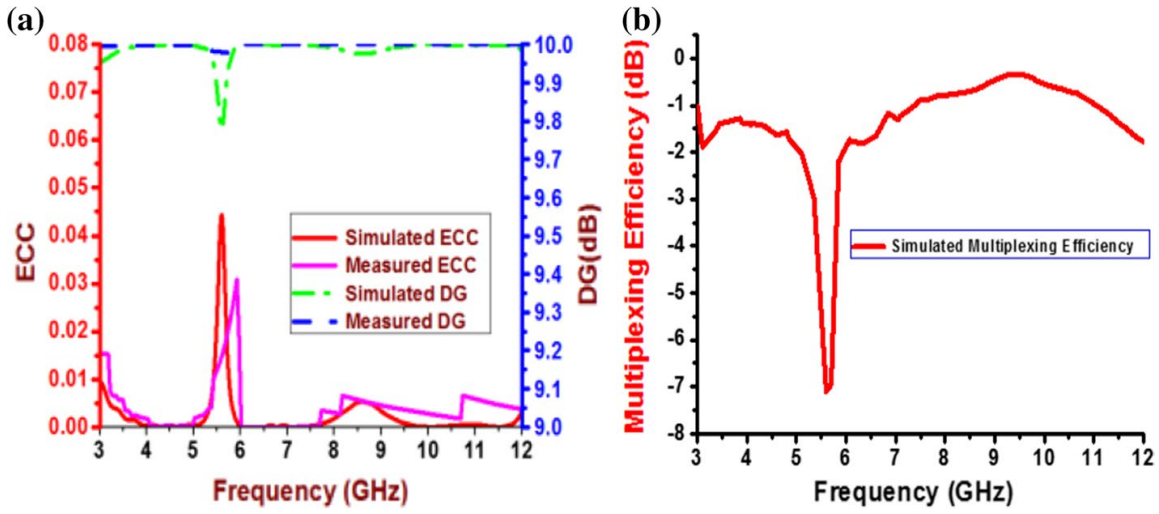


Fig. 11 Simulated and measured results of the proposed antenna **a** envelope correlation coefficient and diversity gain, **b** simulated result of multiplexing efficiency

The result shows that the total efficiency (shown in Fig. 7b) and multiplexing efficiency are almost identical and its value goes downward at notched WLAN band.

The two-port antenna system, contiguous antenna elements impinge on each other and when working concurrently they influence the overall desirable gain, efficiency, and bandwidth. Actual MIMO antenna system performance will not be predicted by S-parameters only, therefore TARC has initiated this effect into account. It is the square root of the ratio of total reflected to total incident power and hence, it gives the information about the apparent return loss of the MIMO antenna system. For side by side two port radiated antenna system, it can be evaluated by Eq. (4).

$$\mathrm{TARC} = \sqrt{\frac{\left(S_{11} + S_{12}\right)^2 + \left(S_{22} + S_{21}\right)^2}{2}} \tag{4}$$

Ideally, the value of TARC should be less than 0 dB for MIMO antenna based system. Figure 12a shows the simulated and measured TARC of the proposed MIMO antenna and it is observed that the value of TARC is less than −28 dB and −40 dB correspondingly in the entire UWB range except for the notched band, it shows the excellent diversity performance of the designed antenna. With increasing the number of elements, the channel capacity of MIMO system is linearly increased, so the element matrix properly characterizes the performance of a MIMO system as it provides the information about the phase and gain characteristics of the transmitting and receiving elements. However, in the presence of the correlation factor in MIMO antennas, the channel capacity losses are also increased. Therefore, diversity performance of proposed antenna pretentious and channel capacity losses are calculated by the following Eq. (5)

$$\mathrm{CCL(dB)} = -\log_2\left[\det\left(\phi^{\mathrm{R}}\right)\right] \tag{5}$$

where φ^{R} is the 2×2 co-relation matrix in terms of S-parameter $\varphi_{ii} = 1 - \left|S_{ii}\right|^2 - \left|S_{ij}\right|^2$, $\varphi_{ij} = -\left(S_{ii} \times S_{ij} + S_{ji} \times S_{jj}\right)$. Figure 12a shows the CCL losses and it is less than 0.35 bits/s/Hz over the entire bandwidth except for notch band, which offer a better diversity results of the proposed MIMO antenna. Moreover, the mean effective gain

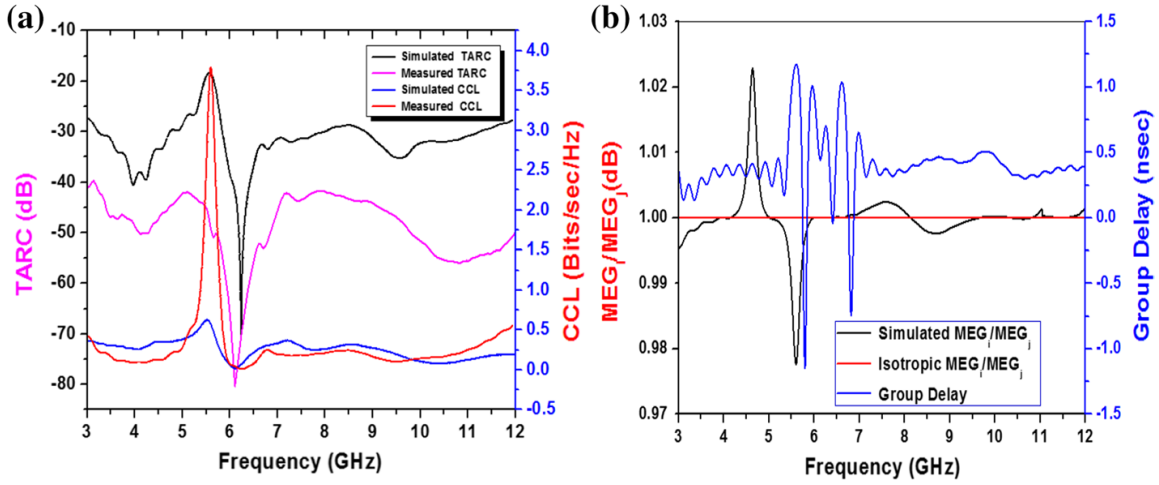


Fig. 12 Simulated and measured results of **a** channel capacity loss (CCL) and total active reflection coefficient (TARC), **b** simulated results of the mean effective gain (MEG) and group delay of proposed MIMO antenna

(MEG) is the ratio of the mean received power by the ith antenna to the mean incident power of the jth antenna with the same route, $MEG_i = \varepsilon_{Total}/2$, where ε_{Total} is implying total effective efficiency of the ith antenna, *i.e.* $\varepsilon^i_{Total} = \varepsilon^i_{mis} \times \varepsilon^i_{rad}$, and where $\varepsilon^i_{mis} = 1 - \sum_{j=1}^{N} \left|S_{ij}\right|^2, \varepsilon E^i_{rad} = \sum_{j=1}^{N} \left|S_{ij}\right|^2$ where ε^i_{mis}, and ε^i_{rad} are mismatch efficiency and radiation efficiency of ith antenna as shown in Fig. 12b. MEG ratio for the proposed MIMO antenna is to be extracted by Eq. (6)

$$MEG_i = \frac{P_{rec}}{P_{inc}} = \oint \left[\frac{XPR \times G_{\theta i}(\Omega) + G_{\phi i}(\Omega) \times P_{\phi}(\Omega)}{1 + XPR} \right] d\Omega \tag{6}$$

where, XPR is the cross-polarization ratio, $G_{\theta i}$ (Ω), $G_{\Phi i}$ (Ω) and $P_{\Phi}(\Omega)$ is the gain and power density function of the incident wave, Ω = beam area.

The acceptable ratio of mean effective gain of ith and jth antenna elements should be less than or equal to ± 3 dB and as from Fig. 12b, the MEG ratio of the projected MIMO antenna found in acceptable limit throughout the UWB range, which is a high-quality channel performance for wireless communication. Finally, to verify the time domain analysis of the intended MIMO antenna, two identical antennas have located in front of each other as a receiver and transmitter at a distance of 100 cm and group delay (ns) is calculated. Group delay is the rate of change of transmission phase angle with respect to frequency and it is almost constant in entire the intended band. Group delay in terms of return loss (S_{11}) is defined by equation $T_d(\omega) = -\delta\phi/\delta\omega$, where Φ = phase of S_{11} and ω = angular frequency in rad/s. The group delay deviation of proposed MIMO antenna is ≤ 0.3 ns entire the UWB frequency span excluding notched band as shown in Fig. 12b.

The MEG is also represented in isotropic and Gaussian mediums, to investigate the diversity performance at XPR = 0 dB (outdoor) and 6 dB (indoor) as shown in Fig. 13a and Table 3 of the proposed antenna. The MEG for the isotropic medium at 0 dB and 6 dB XPR is constant at − 3.0 dB and lies between − 3.1 and − 5.1 dB respectively. Similarly, the MEG for Gaussian medium at 0 dB and 6 dB XPR is lying between − 2.7 and − 4.2 dB, and − 4.6 to − 6.5 dB respectively for entire UWB bandwidth.

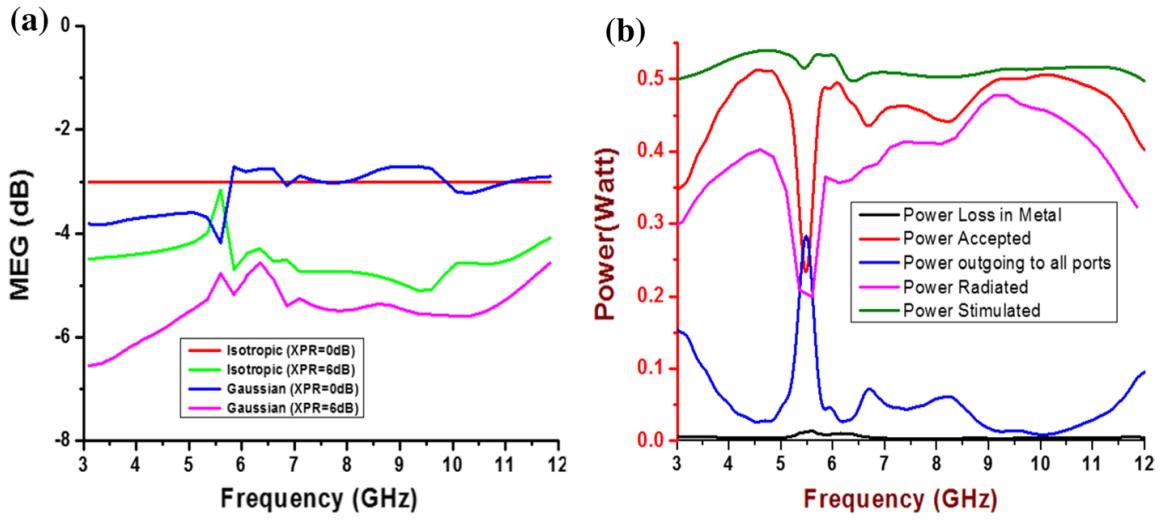


Fig. 13 **a** A simulated mean effective gain in Isotropic and Gaussian medium at various XPR values of proposed antenna applications, **b** power associated with proposed MIMO antenna for various diversity performances

Table 3 Simulated MEG results of proposed MIMO antenna at various frequencies

Frequency (GHz)	MEG (dB) Isotropic medium		MEG (dB) Gaussian medium	
	XPR = 0 dB	XPR = 6 dB	XPR = 0 dB	XPR = 6 dB
4.1	− 3.0	− 4.4	− 3.7	− 6.1
6.6	− 3.0	− 4.5	− 2.8	− 4.9
9.6	− 3.0	− 5.1	− 2.7	− 5.6

The power loss, power accepted, power outgoing, power radiated, and power stimulated of the proposed antenna is used to calculate the parameters like SAR, CCL, ECC, DG and radiation efficiency as shown in Fig. 13b. 0.5 W stimulated power is used to calculate all MIMO antenna parameters. Stimulated power, accepted power, and power radiated is having the same nature with different values due to various losses like a dielectric, conductor, surface wave, and port losses. Therefore, stimulating power is always greater than accepted and radiated power. Power going to all ports is highly interfere at the notched band (5.1–5.85 GHz) so that the maximum power is correlated between ports and not used for radiation.

An evaluation of the proposed MIMO antenna with existing referred MIMO antennas is represented in Table 4. It is observed that the designed antenna achieves the better results in comparison with the most of the MIMO antennas at various specifications, which is highly appreciable as small in size, low in cost, improved radiation efficiency (0.75–0.97) with enhanced isolation, better ECC and diversity gain (> 9.95 dB).

Table 4 Comparison of proposed UWB-MIMO antenna including WLAN notched band with existing references at various aspects

Antenna size (mm^2)	Isolation (dB)	Frequency (GHz)	ECC	DG (dB)	Material	Radiation efficiency
50 × 90 [2]	> 12	1.8–5.2	< 0.33	> 9.50	RO4350B	> 0.63
50 × 40 [3]	> 15	2.5–11	< 0.02	> 9.65	FR-4	> 0.692
50 × 82 [5]	> 15	2.15–13.62	< 0.04	> 9.79	FR-4	0.6-0.85
23 × 18 [6]	> 15	3–12.4	< 0.015	> 9.90	F4b2	> 0.70
36 × 18 [7]	> 20	2.8–20	< 0.03	> 9.95	FR-4	> 0.692
23 × 29.8 [3]	> 21	2.5–12	< 0.01	> 9.95	Rogers TMM4	0.55–0.82
40 × 40 [9]	> 15	2–6	< 0.1	> 9.95	FR-4	0.5–0.85
26 × 15 [11]	> 24	2–35	< 0.03	> 9.90	FR-4	> 0.50
55 × 13.5 [13]	> 18.3	1.92–10.6	< 0.019	> 9.95	FR-4	> 0.91
36 × 22 [Proposed]	> 30	3.1–11.2	< 0.008	> 9.95	FR-4	0.75–0.97

5 Conclusion

A MIMO antenna has been designed for UWB bandwidth (3.1–11.2 GHz) with better isolation (≤ -30 dB), radiation efficiency (> 0.75), and sharp rejection at WLAN (5.1–5.85 GHz) band. Diversity performance of antenna has also been verified in terms of ECC (≤ 0.008), TARC (≤ -25 dB), CCL (≤ 0.3 bits/s/Hz), Mean effective gain ratio (MEG_i) $\cong 1$ and Directive Gain (≥ 9.95 dB). Apart from the above-mentioned antenna performance, the SAR value of antenna, when it is used in human interface devices is less than the permissible limit as mentioned in Table 2, hence the antenna prototype is appropriate for UWB bandwidth at high interference surroundings in indoor or outdoor applications.

References

1. Li, Q., Feresidis, A. P., Mavridou, M., & Hall, P. S. (2015). Miniaturized double-layer EBG structures for broadband mutual coupling reduction between UWB monopoles. *IEEE Transactions on Antennas and Propagation, 63*(3), 1168–1171.
2. Dhar, S. K., Sharawi, M. S., Hammi, O., & Ghannouchi, F. M. (2016). An active integrated ultra-wideband MIMO antenna. *IEEE Transactions on Antennas and Propagation, 64*(4), 1573–1578.
3. Lin, G.-S., Sung, C.-H., Chen, J.-L., Chen, L.-S., & Houng, M.-P. (2017). Isolation improvement in UWB MIMO antenna system using carbon black film. *IEEE Antennas and Wireless Propagation Letters, 16,* 222–225.
4. Mao, C.-X., & Chu, Q.-X. (2014). Compact radiator UWB-MIMO antenna with dual polarization. *IEEE Transactions on Antennas and Propagation, 62*(9), 4474–4480.
5. Toktas, A. (2016). G-shaped band-notched ultra-wideband MIMO antenna system for mobile terminals. *IET Microwaves, Antennas and Propagation, 11*(5), 718–725.
6. Tao, J., & Feng, Q. (2017). Compact ultrawideband MIMO antenna with half-slot structure. *IEEE Antennas and Wireless Propagation Letters, 16,* 792–795.
7. Chandel, R., Gautam, A. K., & Rambabu, K. (2018). Design and packaging of an eye-shaped multiple-input–multiple-output antenna with high isolation for wireless UWB applications. *IEEE Transactions on Components, Packaging, and Manufacturing Technology, 8*(4), 635–642.
8. Khan, M. S., Capobianco, A.-D., Naqvi, A., Ijaz, B., Asif, S., & Braaten, B. D. (2015). Planar, compact ultra-wideband polarisation diversity antenna array. *IET Microwaves, Antennas and Propagation, 9*(15), 1761–1768.
9. Sarkar, D., & Srivastava, K. V. (2017). A compact four-element MIMO/diversity antenna with enhanced bandwidth. *IEEE Antennas and Wireless Propagation Letters, 16,* 2469–2472.
10. Liu, L., Cheung, S. W., & Yuk, T. I. (2015). Compact MIMO antenna for portable UWB applications with band-notched characteristic. *IEEE Transactions on Antennas and Propagation, 63*(5), 1917–1924.
11. Gautam, A. K., Yadav, S., & Rambabu, K. (2018). Design of ultra-compact UWB antenna with band-notched characteristics for MIMO applications. *IET Microwaves, Antennas and Propagation, 12*(12), 1895–1900.
12. Lee, J.-M., Kim, K.-B., Ryu, H.-K., & Woo, J.-M. (2012). A compact ultrawideband MIMO antenna with WLAN band-rejected operation for mobile devices. *IEEE Antennas and Wireless Propagation Letters, 11,* 990–993.
13. Rajkumar, S., Selvan, K. T., & Rao, P. H. (2018). Compact 4 element S Sierpinski K nopp fractal UWB MIMO antenna with dual-band notch. *Microwave and Optical Technology Letters, 60*(4), 1023–1030.
14. Gao, P., He, S., Wei, X., Xu, Z., Wang, N., & Zheng, Y. (2014). Compact printed UWB diversity slot antenna with 5.5-GHz band-notched characteristics. *IEEE Antennas and Wireless Propagation Letters, 13,* 376–379.
15. Abedian, M., Rahim, S. K. A., Fumeaux, C., Danesh, S., Lo, Y. C., & Jamaluddin, M. H. (2017). Compact ultrawideband MIMO dielectric resonator antennas with WLAN band rejection. *IET Microwaves, Antennas and Propagation, 11*(11), 1524–1529.

16. Sengar, K., Rani, N., Singhal, A., Sharma, D., Verma, S., & Singh, T. (2014). Study and capacity evaluation of SISO, MISO, and MIMO RF wireless communication systems. ArXiv preprint arXiv :1403.7774.
17. Rajkumar, S., Sivaraman, N. V., Murali, S., & Selvan, K. T. (2017). Heptaband Swastik arm antenna for MIMO applications. *IET Microwaves, Antennas and Propagation, 11*(9), 1255–1261.
18. Liu, Guifeng., Liu, Ying., & Gong, Shuxi. (2017). Compact uniplanar UWB MIMO antenna with band-notched characteristic. *Microwave and Optical Technology Letters, 59*(9), 2207–2212.
19. Deng, J. Y., Wang, Z. J., Li, J. Y., & Guo, L. X. (2018). A dual-band MIMO antenna decoupled by a meandering line resonator for WLAN applications. *Microwave and Optical Technology Letters, 60*(3), 759–765.
20. Fang, H.-S., Wu, C.-Y., Sun, J.-S., & Huang, J.-T. (2017). Design of a compact MIMO antenna with pattern diversity for WLAN application. *Microwave and Optical Technology Letters, 59*(7), 1692–1697.
21. Ali, W. A., & Ibrahim, A. A. (2017). A compact double-sided MIMO antenna with improved isolation for UWB applications. *AEU-International Journal of Electronics and Communications, 82,* 7–13.
22. Ekrami, H., & Jam, S. (2018). A compact triple-band dual-element MIMO antenna with high port-to-port isolation for wireless applications. *AEU-International Journal of Electronics and Communications, 1*(96), 219–227.
23. Awasthi, Y. K., Singh, H., Sharma, M., Kumari, S., & Verma, A. K. (2017). Computer-aided design-based circuit model of microstrip line for terahertz interconnects technology. *The Journal of Engineering, 9*(7), 512–526.
24. Abd-Alhameed, R. A., Mangoud, M., Excell, P. S., & Khalil, K. (2005). Investigations of polarization purity and specific absorption rate for two dual-band antennas for satellite-mobile handsets. *IEEE Transactions on Antennas and Propagation, 53*(6), 2108–2110.
25. Iyama, T., Onishi, T., Tarusawa, Y., Uebayashi, S., & Nojima, T. (2008). Novel specific absorption rate (SAR) measurement method using a flat solid phantom. *IEEE Transactions on Electromagnetic Compatibility, 50*(1), 43–51.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Gaurav Saxena received his B. Tech degree in Electronics and Communication Engineering, from Uttar Pradesh Technical University Lucknow, India, in 2007, M. Tech degree in Microwave Electronics, from University of Delhi, South Campus, Delhi, India, in 2012. Currently, he is working an Assistant Professor in Galgotia College of Engineering and Technology, Greater Noida, Uttar Pradesh and Pursuing Ph.D. from Delhi Technological University, Delhi, India. His area of research interest are MIMO Antenna design for 5G wireless applications, Metamaterial absorber, and Microwave passive/active components.



Priyanka Jain received her B.E. degree in Electronics and Communication Engineering, M. Tech degree in Microwave Electronics, from the University Of Delhi, South Campus, Delhi, India, and Ph.D. from GGSIPU Delhi, India. Currently, she is working an Assistant Professor, Delhi Technological University, Delhi, India. She has published various articles in peer-reviewed journals and conferences. Her recent research interest includes Modeling of passive microwave components, Microwave antenna for 5G applications, Meta-material absorber, MIMO receiver/transmitter system, etc.



Yogendra Kumar Awasthi (M'13) received the B.Sc (Electronics) degree from Agra University, Agra, India in 1999, the M.Sc (Electronics and Computational Physics) degree from Agra University, Agra, India in 2001, in Electronics and Computational Physics and Ph.D. (RF and Microwave Engineering) from Delhi University, Delhi, in 2012. He was co-operative faculty in University of Delhi, New Delhi and he was also visiting faculty at Central Electronics Engineering Research Institute, Pilani, and Rajasthan, India in 2006–2012. He is currently working as Associate Professor with the Department of Electronics and Communication Engineering, Manav Rachna International Institute of Research and Studies, Faridabad, India. He has published more than 70 articles in peer-reviewed journals and conferences and he has also filed four patents and published one book's chapter. His recent research interest includes Modeling of high-frequency passive microwave components, Microwave antenna for LTE and 5G applications, Meta-material absorber, Sensor & Cloaking, Study of transient, RF Power Transfer and Energy Harvesting, MIMO receiver system, Electrical Power System, etc.

Journal of ELECTRICAL ENGINEERING, VOL 70 (2019), NO6, 443–453

sciendo
PAPERS

Human action recognition using descriptor based on selective finite element analysis

Rajiv Kapoor*, Om Mishra*, Madan Mohan Tripathi**

This paper proposes a novel local descriptor evaluated from the Finite Element Analysis for human action recognition. This local descriptor represents the distinctive human poses in the form of the stiffness matrix. This stiffness matrix gives the information of motion as well as shape change of the human body while performing an action. Initially, the human body is represented in the silhouette form. Most prominent points of the silhouette are then selected. This silhouette is discretized into several finite small triangle faces (elements) where the prominent points of the boundaries are the vertices of the triangles. The stiffness matrix of each triangle is then calculated. The feature vector representing the action video frame is constructed by combining all stiffness matrices of all possible triangles. These feature vectors are given to the Radial Basis Function-Support Vector Machine (RBF-SVM) classifier. The proposed method shows its superiority over other existing state-of-the-art methods on the challenging datasets Weizmann, KTH, Ballet, and IXMAS.

K e y w o r d s: finite element analysis (FEA), stiffness matrix, discretization, support vector machine

1 Introduction

Over the last few decades, it has been observed that computers have transformed human life in almost every possible aspect. Along with the latest transformations, video data has become easily accessible and dominant in the present time. Every new reform has enabled hardware devices like mobile phones, tablets, digital cameras to create, store and share videos. The increasing number of accessible videos has also created the need to understand them. This idea has led to an extensive study of videos to recognize the action. Action recognition has its major application in the field of medical, sports and security.

Researchers modeled action features globally as well as locally. Global features rely on the localization of individual whose action is to be perceived. Localization is done through background subtraction or human tracking. The 2D template methods use 2D silhouettes for global representation [1, 2]. At this point Hu moments, Radon transform descriptors are also utilized to signify the activities. Global features can also be represented as space-time volumes. Spatial-temporal volume is made by stacking silhouettes over a given grouping [3, 4]. Global features may also represent motion information with the assistance of optical flow. It doesn't require background subtraction. An optical flow-based strategy where the movement of pixels have been contemplating is also used in global feature extraction [5, 6]. The disadvantage of this method is that it is very sensitive to noise because the motion descriptor can be corrupted due to noise that appeared in a dynamically changing background. Primary inconveniences of global features portrayal are that it unpleasantly depends on precise localization and background subtraction, so it is sensitive to a viewpoint and individual appearance. Likewise, it cannot give motion information of an activity which makes it sub-par compared to similar kinds of activities like jogging and running.

Local feature portrayal is utilized more frequently in recent times. It doesn't require precise restriction and background subtraction and furthermore demonstrates invariance in viewpoint and individual appearance. Local spatiotemporal descriptors are based on the bag-of-words model [7-9]. HOG/HOF, HOG3D, SURF, and MoSIFT are some major descriptors [10-13]. The principal burden of this portrayal is that they can't give basic/shape information. Its aftereffect is that it can give the same component descriptor for various activity classes. Furthermore, in recent years, a paradigm based on deep learning techniques is also very popular in the research community for human action recognition [14-16]. Unlike the handcrafted approaches discussed above the deep learning, the technique is fully automated. The efficiency of deep learning methods depends upon the design of the network. These methods need large datasets and parameters, on account of this, the complexity of the structure increases. Researchers are working on deep learning techniques to overcome this problem.

Silhouette analysis-based methodologies contributed significantly to human action recognition. They are used to find out both global and local features [17- 21]. In [17] and [18] silhouettes analysis is used to extract the global and local features to represent action video. [19] represented the local feature as pose correlogram and extended motion history image is used for global features.

*Department of Electronics & Communication, Delhi Technological University, Bawana Road, Delhi-110042, India, **Department of Electrical Engineering, Delhi Technological University, Bawana Road, Delhi-110042, India, mmtripathi@dce.ac.in

DOI: 10.2478/jee-2019–0077, Print (till 2015) ISSN 1335-3632, On-line ISSN 1339-309X
© This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivs License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

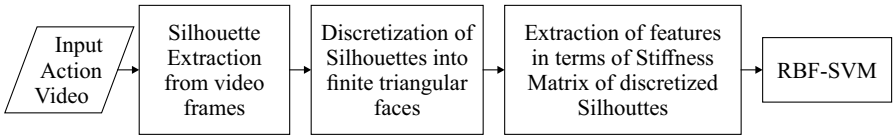


Fig. 1. Workflow diagram of the proposed method

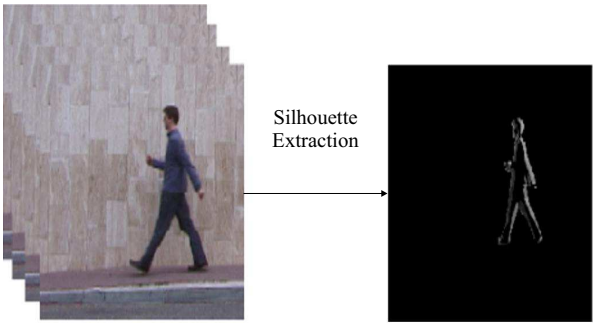


Fig. 2. Walk and the extracted silhouette

The three-dimensional Histogram of the oriented gradient is used to represent the action video in [20]. The method [21] proposed a new feature based on the negative space which is related to the region of the surrounding of the subjects.

The human body poses represent an action effectively. Multi-View key poses from the silhouette are extracted in [22]. They modeled the actions with spatiotemporal features. In [23] the binarized silhouette is used to find out the trace transform to represent the global feature of action The sequence of a silhouette is represented as the cube video to model the action in [24]. The multi-scale volumetric approach for action videos is used in [25, 26]. The action is modeled using sparse coding of image sequences in [26]. The silhouette-based analysis is also used in deep learning-based methodologies [27, 28, and 29]. The CNN and HMM are combined to represent long action video in [27]. The methodology used in [28] represented the actions through a neural mechanism. The two cortical areas, the primary cortex, and the middle cortex are used to extract the motion features. To capture complete motion information, [29] proposed a new descriptor that is capable of static, short term motion as well as long term motions.

Motivation: This paper introduces a new feature descriptor based on Finite Element Analysis (FEA) [30-31]. FEA has been used as a very powerful technique for the structural analysis of the system. In the FEA technique, the structure is converted into a finite number of elements. Wherever any deformation occurs in a body/structure these finite elements also get displaced from their previous position and the stiffness matrix of these elements shows how stiff the body/structure is against this deformation. This gives accurate and precise information about the structural deformation of the body. Similarly, when a person performs an action, his body gets deformed in different patterns. This motivates us to apply the concept of FEA on the silhouettes extracted from the action video. The proposed method offers a new local features descriptor that is solely capable of representing shape as well as motion features of the silhouette.

2 Methodology of proposed framework

The proposed methodology can be described by the Fig. 1. The human silhouette is extracted from the frames of the action video. Then we discretized the silhouette of the human body into several finite elements (triangle faces). Then complete stiffness matrix of the silhouette is calculated by using FEA. The stiffness matrices are represented as feature vectors. The RBF-SVM classifier is used for the classification of actions.

2.1 Silhouette extraction

In the proposed method, features are acquired to distinguish different human actions which make the result more authentic. These features describe the deformation that occurs in the silhouette in terms of shape and motion information while performing an action. As silhouette moves, the finite elements also get deformed. The stiffness matrix of the silhouette narrates these features. The first step of the proposed method is silhouette extraction which is also a very challenging problem because it requires background subtraction. Background cluttering, illumination change, noises, etc. are some challenges for background subtraction. The GMM [32] is robust to problems discussed above and it also has the capability to deal with the critical issue like a shadow. We used GMM for background subtraction. Then the silhouette is extracted and normalized so that all the silhouettes become equal in size [19]. Fig. 2 shows the extracted silhouette from the video.

2.2 Discretization and shape function representation

The preliminary step for FEA is discretization i. e. modeling the silhouette structure into numbers of small elements as shown in Fig. 3(a). The number of elements in which geometry is divided is variable and can be determined by software like MATLAB, COMSOL, *etc.* which demand physics of the geometry. We selected the simple triangular element as the finite element. The reason behind using the triangular shape is that it is the simplest structure for numerical representation. We used MATLAB having an FEA toolbox in the proposed method.

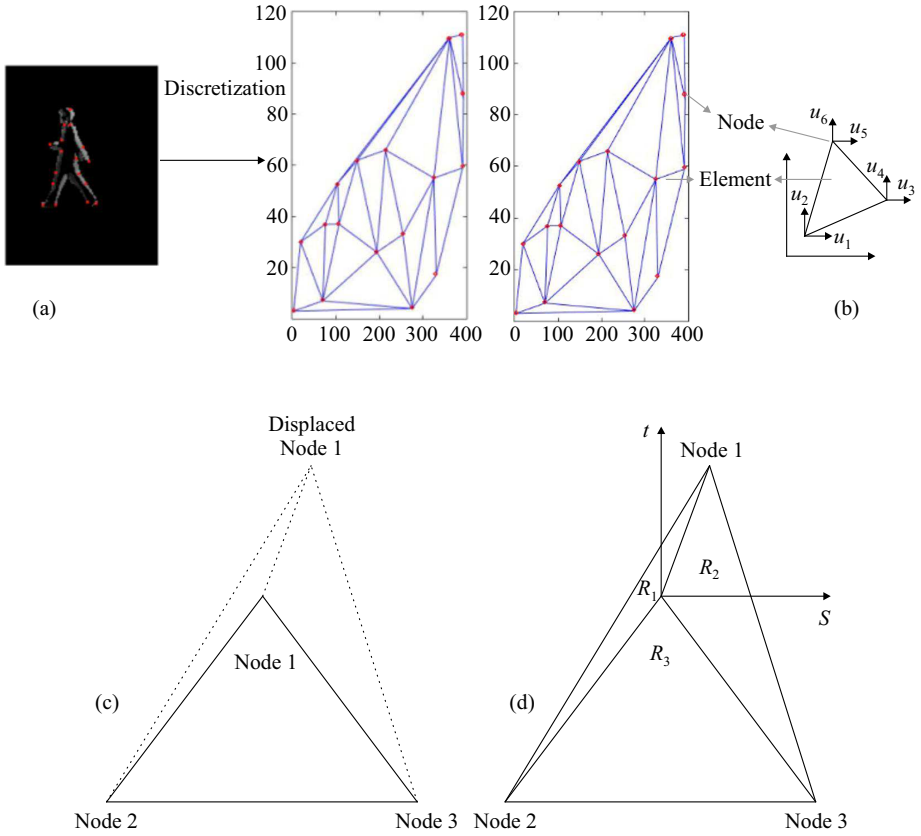


Fig. 3. (a) – segmented discretized silhouette, (b) – displacement of a mesh (triangle) of the human body and its mesh element, (c) – deformed triangle as node 1 is displaced, (d) – triangle divided into 3 parts

We referred Laptev *et al*[10] to find out the prominent points at the boundary of the silhouette. These prominent points are given as the nodes to the FEA toolbox. The silhouette is discretized into the finite triangular elements using these nodes as vertices of the triangle. The discretization of the silhouette is done in such a way that they do not overlap.

In Fig. 3(a) silhouette is discretized into finite triangular elements. The discretized structural representation of the silhouette is shown on the right side of Fig. 3(a) where X -axis and Y -axis are the spatial coordinates of the vertices of the triangle. The complete domain is divided into simpler parts; it provides precise and detailed representation for analysis. Each triangular element has three nodes. Every node has a displacement in the X direction and Y direction as shown in Fig. 3(b). Displacement vectors of the triangle can be represented as

$$U = \{u_1, u_2, u_3, u_4, u_5, u_6\}^\top \tag{1}$$

where, U is a displacement vector for each triangular element and u_1 and u_2 are the displacement of node 1 in X and Y direction respectively, u_3 and u_4 are the displacements of node 2 in X and Y direction and similarly u_5 and u_6 are the displacements of node 3.

When silhouette moves from one frame to another frame, nodes of the triangles are also displaced. To build the face correspondence between the frames we calculated the Euclidean distance among the nodes of the previous frame and next frame. The minimum distance will correspond to the same points. The displacement of the nodes of the triangle is found out from one frame to another frame. Shape function [30,31] of the triangle is used to represent the nodal displacement. Fig. 3(c) shows the displacement of node 1 as a dotted line from its previous point to the displaced point, which results in the deformation of the triangle if the other two nodes are fixed. Similarly, node 2 and node 3 are considered. An interior-point is taken inside the triangle to divide it into 3 regions as shown in Fig. 3(d). Let the total area of the triangle is R and R_1, R_2 and R_3 are the areas of three regions. This is represented by (2).

$$R = R_1 + R_2 + R_3 \tag{2}$$

From (2) the shape functions of all three regions are evaluated using (3)

$$J_1 = \frac{R_1}{R},\ J_2 = \frac{R_2}{R}, \quad \text{and} \quad J_3 = \frac{R3}{R}, \tag{3}$$

where, J_1, J_2, and J_3 are the shape functions of regions R_1, R_2, and R_3. We assume the displacement of the interior point of the triangle in Fig. 3(d) is s and t in X and

Y directions respectively. Displacements s and t can be represented with the help of the shape functions

$$s = J_1 u_1 + J_2 u_3 + J_3 u_5 \tag{4}$$

$$t = J_1 u_2 + J_2 u_4 + J_3 u_6 \tag{5}$$

In (4) and (5), u_1, u_3, and u_5 are the displacement in the X direction and u_2, u_4 and u_6 are the displacements in the Y direction of the vertices of the triangle. The shape functions are determined by the areas of the different regions of the triangle; thus, the regions of the triangle are dependent on each other and can be represented as

$$J_1 + J_2 + J_3 = 1 \tag{6}$$

From (6), we can say that if we know two shape functions, then the third function can be easily calculated

$$J_3 = 1 - J_1 - J_2 \tag{7}$$

Putting these values into (4) and (5) we get the displacement of the interior point of the triangle in terms of s and t in X and Y directions respectively are

$$s = (u_1 - u_5)J_1 + (u_3 - u_5)J_2 + u_5 \tag{8}$$

$$t = (u_2 - u_6)J_1 + (u_4 - u_6)J_2 + u_6 \tag{9}$$

2.3 Representation of feature vector

The displacement of the interior point represents the displacement of the triangle. As discussed above the interior point (x, y) has displacements s in X direction and t in the Y direction. Due to these displacements, a deformation is produced in the triangular element. This deformation is nothing, but the strain developed in the triangular element in X, Y, and shear direction. These are given follows

Strain in X, Y - direction and shear strain are

$$\phi_x = \frac{\partial s}{\partial x} \tag{10}$$

$$\phi_y = \frac{\partial t}{\partial y} \tag{11}$$

$$\phi_x y = \frac{\partial s}{\partial y} + \frac{\partial t}{\partial x} \tag{12}$$

These strains are written in the form of matrices

$$\phi = \begin{bmatrix} \dfrac{\partial s}{\partial x} \\ \dfrac{\partial t}{\partial y} \\ \dfrac{\partial s}{\partial y} + \dfrac{\partial t}{\partial x} \end{bmatrix} \tag{13}$$

Once we get the strain in the triangular element of the discretized silhouette, we found out the stiffness matrix, using FEA [30-31]

$$K_t = C^\top D C t_e R_e \tag{14}$$

where, k_t is the stiffness matrix for a triangle element, C is a displacement matrix subject to strains in X direction, Y direction, and shear strain, t_e is thickness of the body which is constant in case of silhouette, R_e is the area of the triangle and D is a constant matrix

$$D = \frac{\xi}{1-\tau} \begin{bmatrix} 1 & \tau & 0 \\ \tau & 1 & 0 \\ 0 & 0 & \dfrac{1-\tau}{2} \end{bmatrix} \tag{15}$$

where, ξ is Youngs modulus and τ is Poissons ratio and both are constants. We tuned these parameters and we discussed their values in the experimental result section. Further, we converted the stiffness matrix of the triangle k_t into a one-dimensional feature vector [19] by scanning the matrix from the top left to bottom right element by element. The stiffness matrix of the triangle having m rows and m columns will be converted into the one-dimensional feature vector having total $m \times m$ elements. Similarly, we calculated the feature vectors of all possible triangles of the silhouette. The complete stiffness matrix of the silhouette K_s is created by combining all feature vectors of triangles where rows of the matrix represent the triangles associated with the silhouette. To solve the issue that which feature vector of the triangle will be the first row of the Stiffness Matrix of the silhouette we adopted the following strategy:

We scanned the discretized silhouette from top left to bottom right (interior point of the triangle). The first triangle whose interior point is found first in scanning will represent the first row of the matrix K_s and the triangle whose interior point is found last will represent the last row of the K_s matrix. If a silhouette of a frame is discretized into n number of triangle face then the stiffness matrix of silhouette K_s will have n number of rows and $m \times m$ numbers of the column. Further, the complete stiffness matrix of a silhouette is converted into the feature vector with a similar procedure as discussed above. This feature vector represents the frame of the action video.

2.4 Dimension reduction and classification

A frame of an action video at time t is represented by the feature vector extracted from the proposed method. The length of the feature vector of a frame is

$$C = row \times column$$

of the stiffness matrix of the silhouette. Suppose an action sequence consists of S frames, then that action sequence has S feature vectors. This results in a very high dimensional feature space. To reduce the dimensional feature space, we applied Principal component analysis (PCA). Further, these reduced features are given to RBF-SVM classifier [33-34] to recognize the actions. The proposed

methodology can be summarized in the form of the algorithm as follows

Algorithm:

Given an action video, feature vectors can be constructed as follows.

Step 1: Extraction of silhouettes from input video frames.

Step 2: Extraction of prominent points on the boundary of the silhouettes.

Step 3: Prominent points are given as nodes to the FEA toolbox, MATLAB.

Step 4: Silhouettes are discretized into finite triangular elements where nodes act as vertices of the triangle.

Step 5: Each triangular element is represented by three nodes displacement vector (U) of the triangle.

Step 6: Displacement matrix (C) of each triangle is calculated

Step 7: Stiffness matrix (k_t) is calculated for each triangle with the help of displacement matrix C.

Step 8: Complete the Stiffness matrix of the silhouettes is created by combining stiffness matrices (k_t) of all possible triangles of the silhouette.

Step 9: Stiffness matrix of the silhouette is represented as one-dimensional feature vectors

Step 10: Feature vector is calculated for all frames of the action video for all actions.

Step 11: The RBF-SVM Classifier is used for recognition.

3 Experimental result

We have developed our proposed method on MATLAB R2015a. The proposed algorithm has been tested on a system having hardware configuration processor Intel(R) Core (TM) i5-6200U CPU @2.30GHz 2.40 GHz with 8 Gb RAM and 64-bit operating system. To evaluate the performance of the proposed methodology, accuracy is used as the performance parameter in a leave-one-out cross-validation strategy. It can be represented by using a true positive rate (TPR) and a false-positive rate (FPR) represented using true positive (TP), true negative (TN), false posirtie (FP), and false negative (FN)

$$TPR = \frac{TP}{TP + FN} \tag{16}$$

$$FPR = \frac{FP}{FP + TN} \tag{17}$$

where, TPR represents positive cases that are correctly classified and FRP represents negative cases that are incorrectly classified as positive. Accuracy is calculated as

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{18}$$

We have chosen four challenging datasets for action recognition namely Weizmann data set [4], KTH [35], Ballet [36] and IXMAS [37] to evaluate and compare our proposed method. In the Weizmann action dataset, there are ninety videos. The frame rate is 25 frames per second (fps) and resolution is 144180 pixels. It consists of nine different persons who performed a total of ten actions such as running, jumping, waving, bending, *etc.* The sample frame is shown in Fig. 4(a). The KTH dataset comprises six essential exercises, in particular: applauding, waving, boxing, walking, jogging and running. Activities in KTH have been recorded in four different lightings, indoor and outdoor situations and have 100 recordings. But the foundation for all recordings has been kept the same with a static camera with 25 fps and resolution of 160×120 pixels. The states of the recordings in the KTH informational index suffer from camera development and lighting impacts. The sample frame of this dataset is shown in Fig. 4(b).

Ballet is an expressive dance dataset, which comprises of profoundly complex artful dance poses of various on-screen characters. The specimen casings of the dataset are shown in Fig. 4(c). The dataset is acquired from a ballet dance DVD. The foundation in the dataset is basic. Every video grouping comprises of just a single performing artist. The dataset comprises of 44 videos. There are eight different unique activities performed in these videos. IXMAS is a very challenging dataset where 10 distinctive persons are performing every activity three times. These videos have been recorded from different view perspectives where seven different cameras used for recording. These activities incorporate scratching head, looking at the watch, strolling, taking a seat, etc. This dataset offers different challenges by introducing huge appearance change, intra-class varieties, and self-impediments, *etc.* XMAS results have been evaluated for five different camera views. The sample from the IXMAS dataset is shown in Fig. 4(d).

For the parameter settings, we tuned the important parameters on the KTH dataset and similar settings are applied to other datasets. These important parameters are the number of nodes, number of finite elements and feature dimension through PCA. The number of nodes is the prominent points extracted on the silhouette boundary. We have experimented on 5, 10, 15, 17, 20 and 25 prominent points which were considered as nodes. It is clear from Fig. 5.a that when a few prominent points are greater than 15, we achieve a good result. In the proposed method we have taken 17 numbers of nodes because accuracy is varying only 1-2% as we take several points greater than 17.

A next parameter is several finite elements. The number of finite elements has experimented as 10, 15, 20, 22 and 25. Fig. 5(b) shows that the number of elements greater than 20 is giving better accuracy.

The more discretized element we have, the more will be the accuracy of the representation of the body structure. But the tradeoff is that more discretized elements will increase the complexity in terms of time. Thus, we

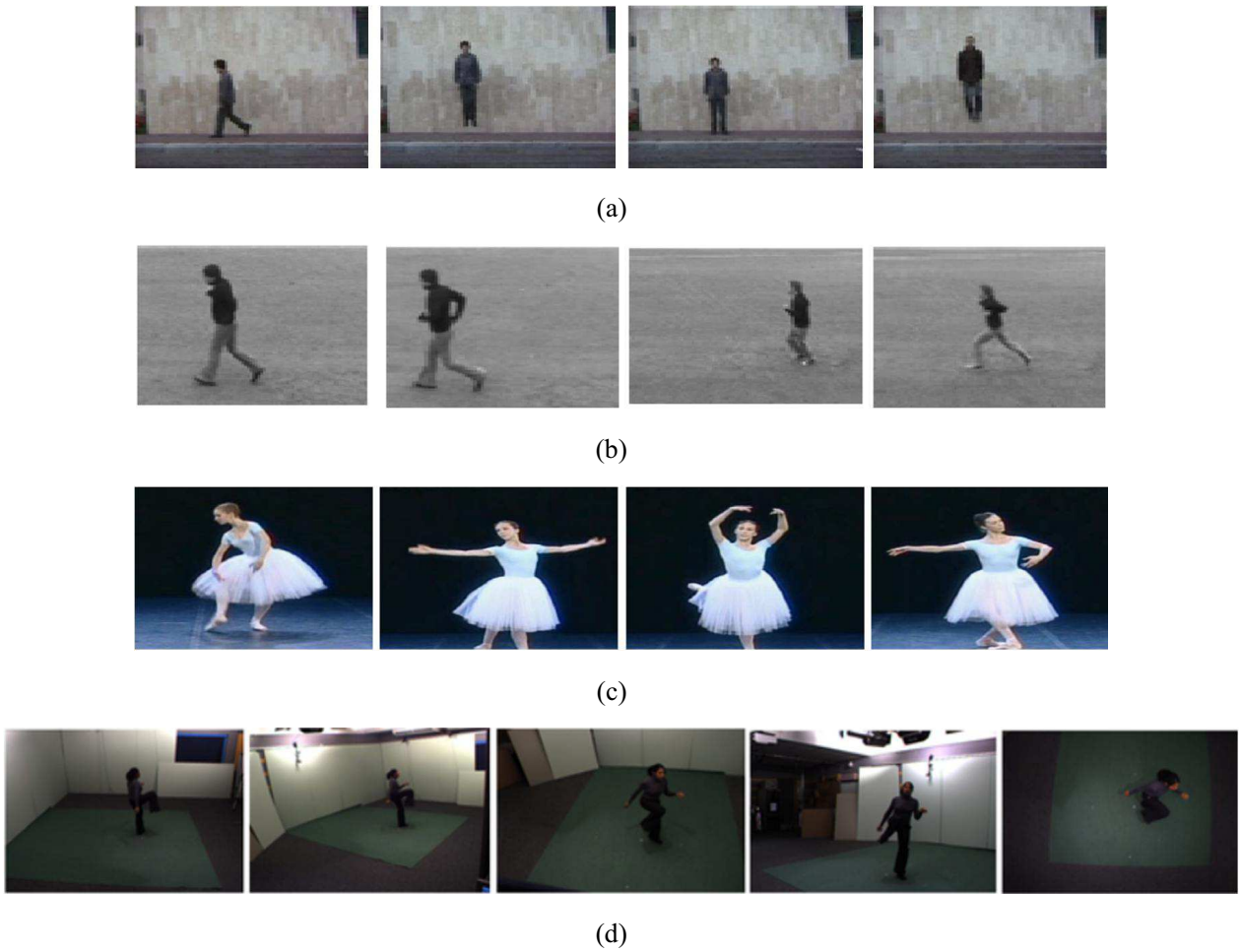


Fig. 4. Datasets:(a) Weizmann,(b) KTH,(c) Ballet,(d)– IXMAS(5 cameras)

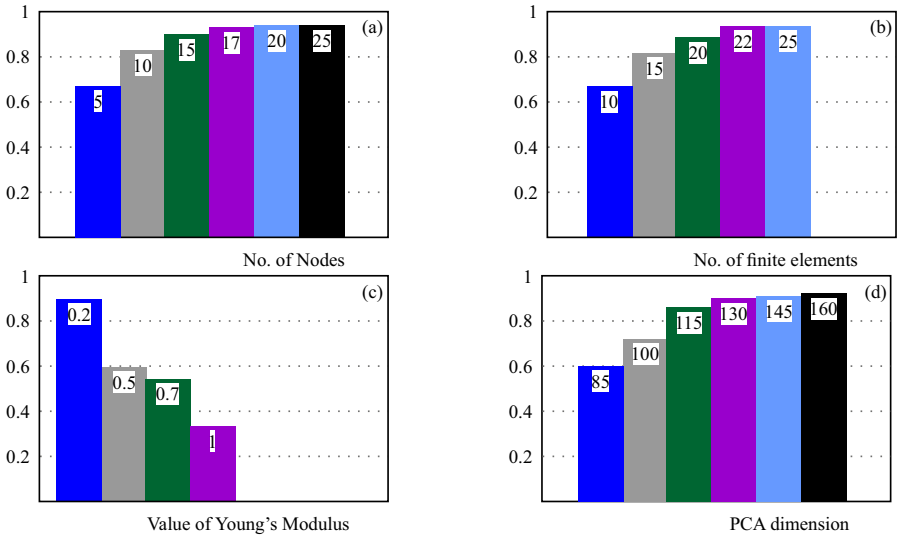


Fig. 5. Parameter setting for: (a) a number of nodes,(b) number of finite elements,(c) Young's Modulus,(d)– feature dimension through PCA

have taken 22 numbers of triangular faces in the proposed method. These 22 numbers of finite elements are taken in such a manner that these triangles do not overlap. This makes the structure simple and attains better accuracy. As far as Young's modulus mentioned in equation (15) is concerned, we have experimented on its normalized values 0. 2, 0. 5, 0. 7 and 1. 0 as shown in Fig. 5(c). We got the highest accuracy when the value of Youngs Modulus was 0. 2. The possible reason behind it could be the value of Youngs modulus is higher for the rigid body and lower

for the flexible body. As the human body is very flexible while performing an action, the lower value 0. 2 gives a better result. The Poissons ratio mentioned in equation (15) is used for the material property and it is a constant

Table 1. Confusion matrix for Weizmann Dataset (R-Running, W-Walking, J-Jumping, JJ-Jumping Jack, S-Skipping, JP-Jumping at a place, SJ-Side Jumping, B-Bending, W-Waving with one hand, WB-Waving with both hands)

	R	W	J	JJ	S	JP	SJ	B	W	WB
R	0.95	0.05	0	0	0	0	0	0	0	0
W	0	1	0	0	0	0	0	0	0	0
J	0	0	1	0	0	0	0	0	0	0
JJ	0	0	0	1	0	0	0	0	0	0
S	0	0	0	0	0.96	0.04	0	0	0	0
JP	0	0	0	0	0.02	0.98	0	0	0	0
SJ	0	0	0	0	0	0	1	0	0	0
B	0	0	0	0	0	0	0	1	0	0
W	0	0	0	0	0	0	0	0	1	0
WB	0	0	0	0	0	0	0	0	0	1

Table 2. Confusion matrix for KTH Dataset (A- Applauding, W-Waving, B- Boxing, WK- Walking, J-Jogging, R-Running)

	A_1	A_2	A_3	A_4	A_5	A_6
A	1	0	0	0	0	0
W	0	1	0	0	0	0
B	0	0.02	0.98	0	0	0
WK	0	0	0	1	0	0
J	0	0	0	0	0.96	0.04
R	0	0	0	0	0.02	0.98

Table 3. Confusion matrix for Ballet LR- Left to right-Hand Opening, RL- Right to left-Hand Opening, J-Jumping, H-Hopping, S-Swinging leg, ST-Standing, T-Turning

	A_1	A_2	A_3	A_4	A_5	A_6	A_7
LR	1	0	0	0	0	0	0
RL	0	1	0	0	0	0	0
J	0	0	0.97	0.03	0	0	0
H	0	0	0.10	0.91	0	0	0
S	0	0	0	0	1	0	0
ST	0	0	0	0	0	1	0
T	0	0	0	0	0	0	1

Table 4. Confusion matrix for IXMAS Dataset: W- Walking, WA- Waving, P- Punching, K- Kicking, T- Throwing, P- Pointing, PU-Picking Up, G- Getting Up, S- Sitting Down, TA-Turning Around, F-Folding arms, C-Checking Watch, SH-Scratching Head

	W	WA	P	K	T	P	PU	G	S	TA	F	C	SH
W	1	0	0	0	0	0	0	0	0	0	0	0	0
WA	0	0.92	0.08	0	0	0	0	0	0	0	0	0	0
P	0	0	0.97	0	0.03	0	0	0	0	0	0	0	0
K	0	0	0	1	0	0	0	0	0	0	0	0	0
T	0	0	0.03	0	0.94	0.03	0	0	0	0	0	0	0
P	0	0	0	0	0.03	0.97	0	0	0	0	0	0	0
PU	0	0	0	0	0	0	0.94	0.06	0	0	0	0	0
G	0	0	0	0	0	0	0.8	0.92	0	0	0	0	0
S	0	0	0	0	0	0	0	0	1	0	0	0	0
TA	0	0	0	0	0	0	0	0	0	1	0	0	0
F	0	0	0	0	0	0	0	0	0	0	0.97	0.03	0
C	0	0	0	0	0	0	0	0	0	0	0.04	0.94	0
SH	0	0	0	0	0	0	0	0	0	0	0	0	1

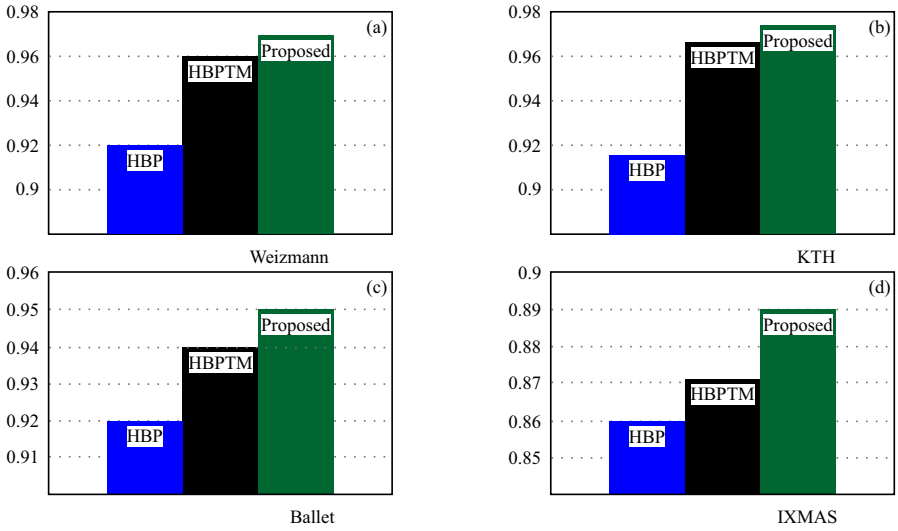


Fig. 6. Comparison of the proposed method with similar methods for four datasets Weizmann, KTH, Ballet, and IXMAS

value that lies between 0-0. 5. In the proposed method we got the optimized result when the value of was 0. 5. The thickness of the silhouette discussed in equation (14) remains constant for all frames in a video and in the proposed methodology; we have taken the value of thickness as 1. The last parameter is the feature dimension through PCA. The result of PCA for different dimensions is represented in Fig. 5(d). We have experimented on different values of dimension such as 85, 100, 115, 130, 145 and 160. Here dimension 130 is showing better results in terms of accuracy and complexity.

We applied the leave-one-out strategy for cross- validation. Tab. 1, Tab. 2, Tab. 3 and Tab. 4 shows the confusion matrices resulted from applying the proposed method on the datasets Weizmann, KTH, Ballet, and IXMAS respectively. These confusion matrices show that most of the actions are 100% classified except some similar types of actions. Thus, we got an accuracy of 97. 8%, 96. 4%, 95. 2% and 90. 3% for the Weizmann, KTH, Ballet, and IXMAS respectively. We compared the proposed method with the other Silhouette analysis based human action recognition methods such as Human Body Pose Model (HBPM) [17, 18, 22] and Human Body Pose Temporal Model (HBPTM) [19, 38] for these datasets. Chaaraoui *et al* . [22] extracted the features of the silhouette as contour points and action is learned from the multi-views of cameras. The multi-view learning makes the method capable of differentiating different persons performing the same action. Wu *et al* . [19] proposed the Human Body Pose Temporal Model where a 2-D silhouette mask is converted into a 1-D feature vector. They represented the action as the correlogram of poses extracted from the silhouette. H. Han *et al* . [25] represented the human body shapes with sparse geometrical features using the Bandlets transformation. They used the AdaBoost to select the features. As the proposed method gives the precise change in the human body shape due to the change in the small elements of the silhouette, it makes it better as compared to other methods. Figure 6 shows that the proposed method shows a better result as compared to other silhouette analysis-based methods using Weizmann, KTH, Ballet, and IXMAS datasets.

We also compared other state-of-the-art methodologies with the proposed method on all four datasets in Tab. 5-Tab. 8. Different testing strategies are used in these methodologies. We mentioned these testing strategies in Tab. s along with the classifier that they have used. Tab. 5 shows a comparison of the proposed method with other methodologies on the Weizmann dataset. Goudelis *et al* . [23] proposed a new feature extraction technique based on the Trace Transform. They represented the spatiotemporal feature in terms of the trace transform from the binarized silhouette. They used the SVM classifier and leave-one-person-out cross-validation testing strategy. They achieved 94.6% accuracy.

The methods [19] and [28] achieved a higher accuracy of 96. 3% and 97. 3% respectively. Liu *et al* . [28] modeled human action with the neural mechanism. They used new feature vectors using two cortical areas one is the primary cortex and the second is the middle temporal cortex for motion. Later, they used the SVM classifier to recognize the action. The proposed method achieved an accuracy of 97. 8%. Since we have discretized the silhouette into a smaller triangle, similar actions such as walking and running are recognizable better than the other methods.

Unlike the Weizmann dataset, the KTH dataset offers more challenging environments. It has different setups having different lighting conditions for different actions. Moreover, the shadow is also a very big challenge in this dataset. So, to deal with these problems, GMM is a better strategy for background subtraction and silhouette extraction. Rahman *et al* . [38] used the negative space-based feature of human pose and motion features to model the actions. To classify the actions, they used Nearest Neighbor Classifier. The leave-one-out strategy is used for cross-validation. They showed an accuracy of

Table 5. Comparison of the proposed method with similar methods on Weizmann dataset

Method	Year	Classifier and Test Scheme	Accuracy
[25]	2015	SVM	81.5
[22]	2013	KNN, LOSO	91.7
[19]	2013	SVM, LOSO	96.3
[23]	2013	SVM, LOPO	94.6
[24]	2014	KNN, LOO	91.4
[26]	2017	NNC	95.3
[27]	2016	CNN-HMM	90.1
[28]	2017	SVM	97.3
Proposed Method		SVM, LOO	97.8

Table 6. Comparison of the proposed method with similar methods on KTH dataset

Method	Year	Classifier and Test Scheme	Accuracy
[25]	2015	Adaboost, SVM, LOO	94.2
[23]	2013	SVM, LOPO	92.7
[26]	2017	NNC, LOO	93.6
[27]	2016	CNN-HMM	94.4
[28]	2017	SVM	91.3
[38]	2014	KNN, LOO	95.1
[39]	2015	KNN	90.8
[29]	2017	CNN-RNN	95.8
Proposed Method		SVM, LOO	96.4

Table 7. Comparison of the proposed method with similar methods on Ballet dataset

Method	Year	Classifier and Test Scheme	Accuracy
[17]	2017	SVM-NN, LOOCV	94.2
[18]	2015	SVM-NN, LOOCV	93.8
[41]	2009	S-CTM, LOO	89.8
[42]	2014	RVM, LOO	90.4
[43]	2014	SVM, LOO	90.3
Proposed Method		SVM, LOOCV	95.2

Table 8. Comparison of the proposed method with similar methods on IXMAS dataset

Method	Year	C1	C2	C3	C4	C5	Overall Accuracy
[44]	2011	89.1	83.4	89.3	87.2	89.2	87.8
[45]	2010	84.2	85.2	84.1	81.5	82.6	82.7
[46]	2013	86.5	83.8	86.1	84.5	87.4	87.2
[47]	2016	91.3	85.7	89.3	90.2	86.5	87.5
Proposed Method		90.8	90 .6	92.4	91.2	90.6	90.2

95. 1% as shown in Tab. 6. In another method, Shi *et al* . [29] proposed new motion descriptor sequential deep trajectory descriptors for long term motion video. The CNN-RNN network is used to learn the motion. They achieved a comparable accuracy of 95. 8% as compared to [38]. We have used a leave-one-out strategy and the proposed method achieved 96. 4% accuracy which is better than other methods.

Table 7 shows a comparison of the proposed method with the other state-of-the-methods for the Ballet dataset. Vishwakarma *et al* . [18] used silhouette-based analysis and extracted the feature vectors based on human poses. They have used SVM, LDA and Neural Network-based hybrid classifiers to recognize the action. They achieved an accuracy of 93. 8%. Vishwakarma *et al* . [17] used a new silhouette analysis where they first found out the average energy image of a silhouette. The spatial distribution of gradient is applied on average energy image to make it a global feature and the temporal feature is found out by Radon transform of the silhouettes. These features are given to the hybrid classifier and they achieve a higher accuracy of 94. 2%. In both methods [17] and [18] they used leave-one-out cross-validation. The proposed method shows better accuracy of 95. 3%. As discussed above discretized silhouette into small triangles helps to recognize the actions in an expressive dataset like Ballet dance. In this dataset, the closely the performer's expressions are observed the better results could be achieved.

Abbreviations used in Tab. 5-Tab. 8 are SVM: support vector machine, KNN: k-nearest neighbor, LOSO: leave-one-sequence -out, LOPO: leave-one-person-out, LOO: leave-one-out, NNC: nearest neighbor classifier, CNN: convolutional neural network, HMM: hidden Markov model, RNN: recurrent neural network, SVM-NN: sup-

port vector machine-neural network, LOOCV: leave-one-out cross-validation, S-CTM, RVM: relevance vector machine.

IXMAS dataset has five different camera views. Methods [42, 46-47] show almost similar accuracies which are around 87%. Wang *et al* . [47] used the Bag-of-visual-word method based on local features. Then they used the cross-view approach to deal with the problem of view change due to different cameras. The proposed method has achieved higher Accuracy for all the views. We got an average of 90. 2% accuracy. To deal with the problem of variation in viewpoint we used view-invariant interest point on the boundary of the silhouette which acted as the vertices of the triangles during the discretization step. Tab. 8 shows a comparison of the proposed method the other state-of-the-methods for IXMAS dataset

For run-time analysis, we have used NVIDIA GPU and MATLAB 2015a with a Parallel computing toolbox. Time taken for several modules in the proposed methodology is calculated. We have analyzed the run-time of the proposed method on all the datasets discussed above. The average run-time of the proposed dataset and is given below stepwise:

Extraction of silhouette from action video (Sec): 0. 41
Silhouette discretization into triangular faces (Sec): 0. 54
Calculation of the silhouette stiffness matrix (Sec): 1. 45
Classification (Sec): 0. 52
Total time for action recognition (Sec): 2.92

Thus, the run-time of the proposed method is fairly good.

6 CONCLUSION

This is a new method to recognize human action through Finite Element Analysis (FEA). A new feature descriptor where the feature vectors of the video frames are expressed in terms of the stiffness matrix of the silhouette extracted from the frames of the video is applied. This offers uniqueness to this method, as it can extract both shapes as well as motion information. The feature vectors extracted from the proposed method are given to the RBF-SVM classifier. Validation of the proposed method has been performed in different challenging environments. The limitation of the methodology is that it requires accurate silhouette extraction. The proposed method shows its superiority as compared to other existing methods of applying them on challenging standard datasets such as Weizmann, KTH, Ballet, and IXMAS.

References

[1] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates, *IEEE Transactions on Pattern Analysis Machine Intelligence*, vol. 23, no. 3, pp. 257-267, 2001.

[2] R. Souvenir and J. Babbs, "Learning the viewpoint manifold for action recognition, *IEEE International Conference on Computer Vision Pattern Recognition (CVPR'08)*, pp. 1-7, 2008.

[3] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Action as space-time shapes, *IEEE International Conference on Computer Vision (ICCV'05)*, vol. 2, pp. 1395-1402, 2005.

[4] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Action as space-time shapes, *IEEE Transaction on Pattern Analysis Machine Intelligence*, vol. 29, no. 12, pp. 2247-2253, 2007.

[5] K. Guo, P. Ishwa, and J. Konrad, "Action recognition from video using feature covariance matrices, *IEEE Transaction on Image Processing*, vol. 22, no. 6, pp. 2479-2494, 2013.

[6] Y. Chen, Z. Li, X. Guo, Y. Zhao, and A. Cai, "A spatio-temporal interest point detector based on vorticity for action recognition, *IEEE International Conference on Multimedia Expo Workshop*, pp. 1-6, 2013.

[7] M. Laptev, C. Marszalek, and B. Schmid, "Learning realistic human actions from movies, *IEEE Conference on Computer Vision Pattern Recognition*, pp. 1-8, 2008.

[8] S. Savarese, A. Delpozo, J. C. Niebles, and L. Fei-fei, "Spatial-temporal correlations for unsupervised action classification, *Proceedings*, of the IEEE Workshop on Motion Video Computing, pp. 1-8, 2008.

[9] M. S. Ryoo and J. K. Aggarwal, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities, *IEEE 12th International Conference on Computer Vision*, pp. 1593-1600, 2009.

[10] I. Laptev and T. Lindeberg "Space-time interest points, *Proceedings Ninth IEEE International Conference on Computer Vision*, pp. 432-439, 2003.

[11] A. Klaser, M. Marszalek and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients, *Proceedings of British Machine Vision Conference*, pp. 995-1004, 2008.

[12] G. Willems, T. Tuytelaars, and L. Van Gool, "An efficient dense scale-invariant spatio-temporal interest point detector, *ECCV 5303*, pp. 650-663, 2008.

[13] M. Chen and A. Hauptmann, "MoSIFT: Recognizing human actions in surveillance videos, *CMU-CS-09-161 2009*,.

[14] N. Ballas, L. Yao, C. Pal, and A. Courville, "Delving deeper into convolutional networks for learning video representations, *International Conference on Learning Representations 2016*,.

[15] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory pooled deep-convolutional descriptors, *IEEE Conference on Computer Vision Pattern Recognition*, pp. 4305-4314, 2015.

[16] L. Sun, K. Jia, D. Yeung, and B. E. Shi, "Human action recognition using factorized spatio-temporal convolutional networks, *IEEE International Conference on Computer Vision (ICCV)*, pp. 4597-4605, 2015.

[17] D. K. Vishwakarma and K. Singh, "Human activity recognition based on the spatial distribution of gradients at sub-levels of average energy silhouette images, *IEEE Transactions on Cognitive Development Systems*, vol. 9, no. 4, pp. 316-327, 2017.

[18] D. K. Vishwakarma and R. Kapoor, "Hybrid classifier based human activity recognition using the silhouettes ands cells, *Expert Systems with Applications*, vol. 42, no. 20, pp. 6957-6965, 2015.

[19] D. Wu and L. Shao, "Silhouette analysis-based action recognition via exploiting human poses, *IEEE Transactions on Circuits Systems for Video Technology*, vol. 23, no. 2, pp. 236-243, 2013.

[20] D. Weinland, M. Ozuysal, and P. Fua, "Making action recognition robust to occlusions viewpoint changes," *European Conference on Computer Vision* (ECCV), pp. 635-648, 2010.

[21] B. Saghafi and D. Rajan, "Human action recognition using Pose-based discriminant embedding, *Signal Processing: Image Communication*, vol. 27, no. 1, pp. 96-111, 2012.

[22] A. A. Chaaraoui, P. C. Pérez, and F. Florez-Revuelta, "Silhouette-based human action recognition using sequences of key poses, *Pattern Recognition Letters*, vol. 34, no. 15, pp. 1799 -1807, 2013.

[23] G. Goudelis, K. Karpouzis, and S. Kollias, "Exploring trace transform for robust human action recognition, *Pattern Recognition*, vol. 46, no. 12, pp. 3238-3248, 2013.

[24] R. Touati and M. Mignotte, "MDS-based multi-axial dimensionality reduction model for human action recognition, *Canadian Conference on Computer Robot Vision*, pp. 262-267, 2014.

[25] H. Han and X. J. Li, "Human action recognition with sparse geometric features, *The Imaging Science Journal*, vol. 63, no. 1, pp. 45-53, 2015.

[26] Y. Fu, T. Zhang, and W. Wang, "Sparse coding-based space-time video representation for action recognition, *Multimedia Tools Applications*, vol. 76, no. 10, pp. 12645-12658, 2017.

[27] J. Lei, G. Li, J. Zhang, Q. Guo, and D. Tu, "Continuous action segmentation recognition using hybrid convolutional neural network-hidden Markov model, *IET Computer Vision*, vol. 10, no. 6, pp. 537-544, 2016.

[28] H. Liu, N. Shu, Q. Tang, and W. Zhang, "Computational model based on the neural network of visual cortex for human action recognition, *IEEE Transactions on Neural Networks Learning Systems*, vol. 29, no. 5, pp. 1427-1440, 2017.

[29] Y. Shi, Y. Tian, Y. Wang, and T. Huang, "Sequential deep trajectory descriptor for action recognition with threestream CNN, *IEEE Transactions on Multimedia*, vol. 19, no. 7, pp. 1510-1520, 2017.

[30] 2D Triangular Elements, The University of New Mexico, http: //www.unm. edu/ bgreen/ME360/2D%20Triangular %20 Elements.pdf. Accessed 24 February 2010,.

[31] D. K. Jha, T. Kant, and R. K. Singh, "An accurate two dimensional theory for deformation stress analysis of functionally graded thick plates, *International Journal of Advanced Structural Engineering*, pp. 6-7, 2014.

[32] J. Dou and J. Li, "Robust human action recognition based on spatiotemporal descriptors motion temporal templates, *Optik*, vol. 125, no. 7, pp. 1891-1896, 2014.

[33] Q. Song, W. Hu, and X. Wenfang, "Robust support vector machine for bullet hole image classification, *IEEE Transaction on Systems Man Cybernetics*,, vol. 32no. pp. 440-448, 2002.

[34] S. S. Keerthi C.-J. Lin, "Asymptotic Behaviors of Support Vector Machines with Gaussian Kernel, *Neural Computation vol*, 15, no, 7,, pp. 1667-1689, 2003.

[35] C. Schuldt, I. Laptev, and B. Caputo, "R, *ognizing human actions: a local SVM approach*, Proceedings of the 17th International Conference on Pattern Recognition Cambridge, UK, 2004,.

[36] T. Guha and R. K. Ward, "Learning sparse representations for human action recognition, *IEEE Transaction on Pattern Analysis Machine Intelligence*, vol. 34, no. 8, pp. 1576-1588, 2012.

[37] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history vol. s, *Computer Vision Image Understanding*, vol. 104, no. 2-3, pp. 249-257, 2006.

[38] S. A. Rahman, I. Song, M. K. H. Leung, I. Lee, and K. Lee, "Fast action recognition using negative space features, *Expert Systems Applications*, vol. 41, no. 2, pp. 574-587, 2014.

[39] I. Gomez-Conde and D. N. Olivieri, "A KPCA spatio-temporal differential geometric trajectory cloud classifier for recognizing human actions in a CBVR system, *Expert Systems Applications*, vol. 42, no. 13, pp. 5472-5490, 2015.

[40] L. Juan and O. Gwun, "A comparison of SIFT, PCA-SIFT and SURF, *International Journal of Image Processing*, vol. 3, no. 4, pp. 143-152, 2009.

[41] Y. Wang and G. Mori, "Human action recognition using semilatent topic models, *IEEE Transactions on Pattern Analysis Machine Intelligence*, vol. 31, no. 10, pp. 1762-1764, 2009.

[42] L.-M. Xia J.-X. Huang, and L.-Z. Tan, "Human action recognition based on chaotic invariants, *Journal of Central University*, vol. 20, no. 11, pp. 3171-3179, 2014.

[43] A. Iosifidis A Tefas and I. Pitas, *Discriminant bag of words based representation for human action recognition*, Pattern Recognition Letters, vol. 49, no. 1, pp. 185-192, 2014.

[44] X. Wu, D. Xu, L. Duan, and J. Luo, "Action recognition using context appearance distribution features, *IEEE Conference on Computer Vision Pattern Recognition (CVPR)*, pp. 489-496, 2011.

[45] D. Weinland, M. Özuysal, and P. Fu, "Making action recognition robust to occlusions viewpoint changes", *European Conference on Computer Vision* (ECCV), pp. 635-648, 2010.

[46] E.-A, Mosabbeb, K. Raahemifar, and M. Fathy, "Multi-view human activity recognition in distributed camera sensor networks, *Sensors*, vol. 13, no. 7, pp. 8750-8770, 2013.

[47] J. Wang, H. Zheng, J. Gao, and J. Cen, "Cross-view action recognition based on a statistical translation framework, *IEEE Transactions on Circuits Systems for Video Technology*, vol. 26, no. 8, pp. 1461-1475, 2016.

Received 18 October 2019

Rajiv Kapoor (Dr) is a Professor in Delhi Technological University, Delhi, India. He worked as Principal in AIACTR, Delhi, India in diverted capacity. He is PhD in Electronics and Communication Engineering from Punjab Engineering College, India. His Research interest include vision/speech-based tracking, activity recognition vision/speech based, signal processing, pattern recognition. He has published more than 100 research articles in leading journals, conference proceedings and books including IEEE, Springer, Elsevier, *etc*

Om Mishra is a research scholar in Department of Electronics & Communication, Delhi Technological University, Delhi, India. He has worked as an Assistant Professor in GB Pant Government Engineering College, New Delhi, India. He received Master of Engineering in Electronics & Communication Engineering from Delhi College of Engineering (Presently DTU), Delhi, India. His research interest includes vision-based activity recognition, signal processing, pattern recognition.

Madan Mohan Tripathi (Dr) is a Professor in Electrical Engineering Department of Delhi Technological University, Delhi, India. He has also worked as Scientist with the Institute for Plasma Research, India and National Institute of Electronics & Information Technology, India. He is PhD in Electrical Engineering from GB Technical University, India. His research interests include Artificial Intelligence applications, renewable energy and power system restructuring. He has published approximately 100 research articles in leading journals and conference proceedings including IEEE, Springer, Elsevier, *etc*

Materials Research Express

ACCEPTED MANUSCRIPT

Investigations on weld bead geometry and microstructure in CMT, MIG pulse synergic and MIG welding of AA6061-T6

To cite this article before publication: Yashwant Koli *et al* 2019 *Mater. Res. Express* in press https://doi.org/10.1088/2053-1591/ab61b6

Manuscript version: Accepted Manuscript

Accepted Manuscript is "the version of the article accepted for publication including all changes made as a result of the peer review process, and which may also include the addition to the article by IOP Publishing of a header, an article ID, a cover sheet and/or an 'Accepted Manuscript' watermark, but excluding any other editing, typesetting or other changes made by IOP Publishing and/or its licensors"

This Accepted Manuscript is **© 2019 IOP Publishing Ltd**.

During the embargo period (the 12 month period from the publication of the Version of Record of this article), the Accepted Manuscript is fully protected by copyright and cannot be reused or reposted elsewhere.
As the Version of Record of this article is going to be / has been published on a subscription basis, this Accepted Manuscript is available for reuse under a CC BY-NC-ND 3.0 licence after the 12 month embargo period.

After the embargo period, everyone is permitted to use copy and redistribute this article for non-commercial purposes only, provided that they adhere to all the terms of the licence https://creativecommons.org/licences/by-nc-nd/3.0

Although reasonable endeavours have been taken to obtain all necessary permissions from third parties to include their copyrighted content within this article, their full citation and copyright line may not be present in this Accepted Manuscript version. Before using any content from this article, please refer to the Version of Record on IOPscience once published for full citation and copyright details, as permissions will likely be required. All third party content is fully copyright protected, unless specifically stated otherwise in the figure caption in the Version of Record.

View the article online for updates and enhancements.

This content was downloaded from IP address 14.139.251.100 on 19/12/2019 at 04:48

Investigations on weld bead geometry and microstructure in CMT, MIG pulse synergic and MIG welding of AA6061-T6

Yashwant Koli[1], N.Yuvaraj[2], Vipin[3], S.Aravindan[4]
[1,2,3]Department of Mechanical Engineering, Delhi Technological University, Delhi 110042, India
[4]Department of Mechanical Engineering, IIT Delhi, New Delhi 110016, India
yashwantkoli5@gmail.com https://orcid.org/0000-0002-3000-8244
yuvraj@dce.ac.in
vipin@dce.ac.in
aravindan@mech.iitd.ac.in

Abstract
This research paper focuses on a comparative study on weld bead geometries of three different welding techniques: Cold Metal Transfer (CMT), Metal Inert Gas Pulse Synergic (MIG P) and MIG Manual Standard (MIG M). Bead-on-plate tests were performed using ER4043 (AlSi5%) as a filler material on the 3.18 mm thick plates of AA6061-T6. Current (80 A, 100 A and 120 A) and welding speed (7.5, 10.5 and 13.5 mm/sec) were used as input process parameters while shielding gas flow rate and contact tip to workpiece distance (CTWD) were maintained constant as 15 l/min and 10mm respectively. The weld beads processed by all the three techniques are compared by analysing the weld bead geometry. Microstructural characterization is carried out using optical microscopy and Field Emission Scanning Electron Microscope (FESEM). CMT has high dilution and penetration with low heat input. Compared to MIG Pulse Synergic and MIG Manual, CMT shows a drastic reduction in residual stresses. A good amount of penetration and dilution with low heat input is necessitated for better joint efficiency.

Keywords: Cold Metal Transfer, MIG pulse synergic, MIG manual, weld bead geometry, dilution, heat input and penetration.

1. Introduction

In the present scenario, industries are moving towards a lighter material, which can significantly reduce the cost of transportation or energy consumption during transportation. Aluminium alloys have eventually managed to replace steel in most of the applications, such as aeronautical and aerospace components, automotive vehicle body parts, external and internal body panels, railcars, shipbuilding, boiler manufacturing, structural and weldable components, etc., due to their high strength, low weight and excellent corrosion resistance [1-8]. AA6061-T6 consists of Mg and Si as the most important alloying elements. A high strength Al-Mg(0.84 %)-Si(0.665 %) alloy contains 0.067 % of manganese to increase ductility and toughness. AA6061-T6, which is a heat treatable alloy being obtained through artificial aging at a low temperature of approximately 180 ºC for 2 hours until it reaches a stable condition. This increases strength to a greater level after solution heat-treating quicker than natural aging.

Cold Metal Transfer (CMT) is an upgraded technology of Gas Metal Arc Welding (GMAW) method, which is based on conventional short-circuiting (CSC) transfer process established via “Fronius of Austria” [9-12]. In CMT welding, 1/4th of the total time will be in a short-circuit phase where the magnitude of current tends to almost zero. It greatly reduces the cost of welding and energy consumption by 30-40% [8]. It is unlike from the conventional short-circuiting (CSC) method given by GMAW. The major difference between CSC and CMT is the mode of liquid droplet detachment from the filler wire. In CSC, Lorentz forces are responsible for droplet detachment whereas in CMT, wire retraction motion is responsible. CMT welding gives the flexibility to weld thin sheets without deformation because of low thermal heat input due to very low non-zero current [13]. It is possible to weld extremely thin sheets like aluminum foil (0.3mm) using CMT. It has a high gap bridging ability and provides spatter-free welds.

Weld parameters directly influence the bead dimensions that in turn dictate the quality of weld [14, 15]. There are many factors that can be taken as welding process parameters like current (I), voltage (V), wire feed rate (WFR), welding speed

Accepted Manuscript

(S), contact tip to workpiece distance (CTWD), stick-out distance and flow rate of shielding gas. From an extensive literature survey and preliminary tests conducted, the notable thing is that "I" and "S" are the most dominant factors that influence bead dimensions and shape relationships [16]. Current, voltage and WFR are the dependent process parameters in CMT and MIG pulse synergic (MIG P). Figure 1 demonstrates the schematic diagram of weld geometry. In this, the shaded portion "A" is reinforcement area and "B" is penetrated area. Dilution is the ratio of molten base metal (BM) volume/area to the volume/area of the total fusion zone (FZ). It can also expressed as the ratio of penetrated area (B) to the area of total weld metal (A+B) as shown in equation 1. The absolute importance of dilution is that the final weld metal's chemical composition is totally distinct from the substrate and filler components, creating a fresh intermediate alloy between the two [17-19]. Dilution enhances the mechanical properties of the weld bead by introducing a suitable filler material to the BM. Like in chemistry, it is stated that dilution is the method of reducing the concentration of solute in a solution by merely adding some amount of solvent like adding water to a solution. Similarly, in welding, some amount of filler wire is fused to the BM, which is calculated by equation 1.

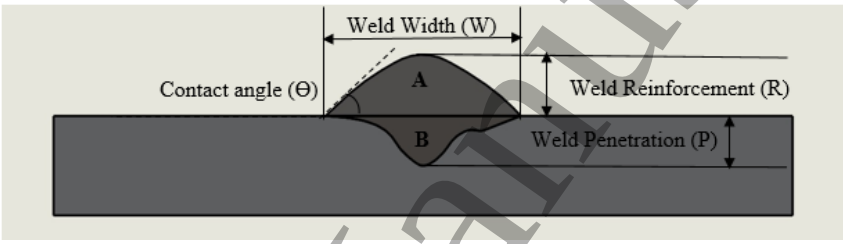


Figure 1. Schematic diagram of weld bead dimensions

Dilution is expressed as the fraction of the parent metal in the subsequent weld metal (WM) and, for a solitary weldment deposit [20]. Analysis of the weld parameters by grey relational analysis (GRA) on MIG welding of high-strength stainless steel determined that welding current had a major effect on the fusing process. Current, voltage and flow rate of gas are 130 A, 27 V and 17 L/min respectively with a 0.8 mm electrode wire diameter are optimum parameters [21]. Welding on AA2012 plate which indicates the impact of current (I), voltage (V) and flow rate of shielding gas increase the weld penetration (P), weld width (W), and reduces the height of weld reinforcement (R) [22]. Impact of welding voltage (V), welding current (I), flow rate of gas and welding speed (S) on tensile strength of st-37 low alloy steel material on MIG welding. The tests use L9 orthogonal array, and the findings showed that current and voltage have an important impact on UTS [23]. Development of mathematical model in MIG welding of SS409M for predicting the weld bead dimensions using central composite rotatable design technique. It is found to be a powerful tool for finding optimal results with lesser number of trials [24, 25]. The ANN indicates the combination of processing parameters that lead to appropriate or better joints while working under different conditions. It makes it possible to trace defects and the quality index maps in the whole domain [26].

Results showed that all bead dimensions are increasing by raising the current and reducing the welding speed. Weld reinforcement form factor (WRFF) is the ratio of bead width (W) to reinforcement height (R) and calculated by equation 2. Weld penetration shape factor (WPSF) is ratio of bead width (W) to penetration (P) which is calculated by equation 3. Heat input is very important term in context to the material thickness and is calculated using equation 4. It is a comparative measure of the transfer of energy per unit weld length. A more effective method needs less thermal input for the same joint, ensuing in stronger weld bead [27]. Heat input has a remarkable influence on the shape factors such as bead width, WRFF and WPSF [19].

Accepted Manuscript

$$\text{Dilution (D)} = \frac{\text{Area B}}{\text{Area A+Area B}} \text{x100} \quad (1)$$

$$\text{WRFF} = \frac{W}{R} \quad (2)$$

$$\text{WPSF} = \frac{W}{P} \quad (3)$$

$$\text{Heat Input (Q)} = \eta \frac{VI}{S} \quad (4)$$

Where Q, V, I, S, η are heat input (J/mm), voltage (V), current (A), welding speed (mm/sec) and efficiency (%) respectively. The thermal efficiency for welding is approximately 80 %.

CMT being a low heat input arc welding technique that welds thin plates with lesser energy consumption. The weld bead dimensions between CMT, MIG P and MIG manual (MIG M) welding techniques are to be compared. The variation in penetration, dilution, heat input, residual stress and microhardness between these three welding techniques are analyzed. Microstructural characterization of these three welding techniques at the bead and fusion zone were scrutinized.

2. Experiment

2.1.Material and methods

AA6061-T6 is chosen as a substrate material of thickness 3.18 mm and ER4043 (AlSi5 %) is used as the filler material of 1.2 mm diameter. Chemical compositions of substrate materials and filler material (wt.%). are tested by chemical spectroscopy method as per ASTM E 1251:2011 and the same is presented in table 1. Mechanical tests were performed on substrate material (AA6061-T6). Microstructural images captured by optical microscope at 200X and 500X magnification are shown in figure 2(a & b). Tensile strength of 284 MPa and total elongation of 24.5 % is clearly shown in figure 2(c). Average microhardness of 116.1 HV is measured by Vicker's microhardness. Substrate is having 97.76 GPa modulus of elasticity. Brinell hardness is measured with a 5 mm diamond indenter giving 93 HB of hardness. XRD plot for substrate material is shown in figure 2(d). It clearly shows the primary elements (Al and Si) at various peaks with varying intensity, which constitutes the material. Figure 2(e) shows the histogram plot for grain size of parent material with an average grain size of 28 µm.

Table 1. Chemical compositions (wt. %) of parent materials and filler metal as per ASTM E 1251:2011

Composition	Al	Si	Fe	Zn	Cu	Ti	Mg	Mn	Cr
AA6061-T6	97.75	0.665	0.253	0.0226	0.153	0.0225	0.840	0.0676	0.178
ER4043	Bal.	5.6	0.8	0.10	0.3	0.02	0.05	0.05	0.05

Table 2. Process parameters for CMT, MIG P and MIG M with their levels

Welding Process Parameters	Units	Symbol	Levels		
			-1	0	1
Current	A	I	80	100	120
Welding speed	mm/sec	S	7.5	10.5	13.5

Accepted Manuscript

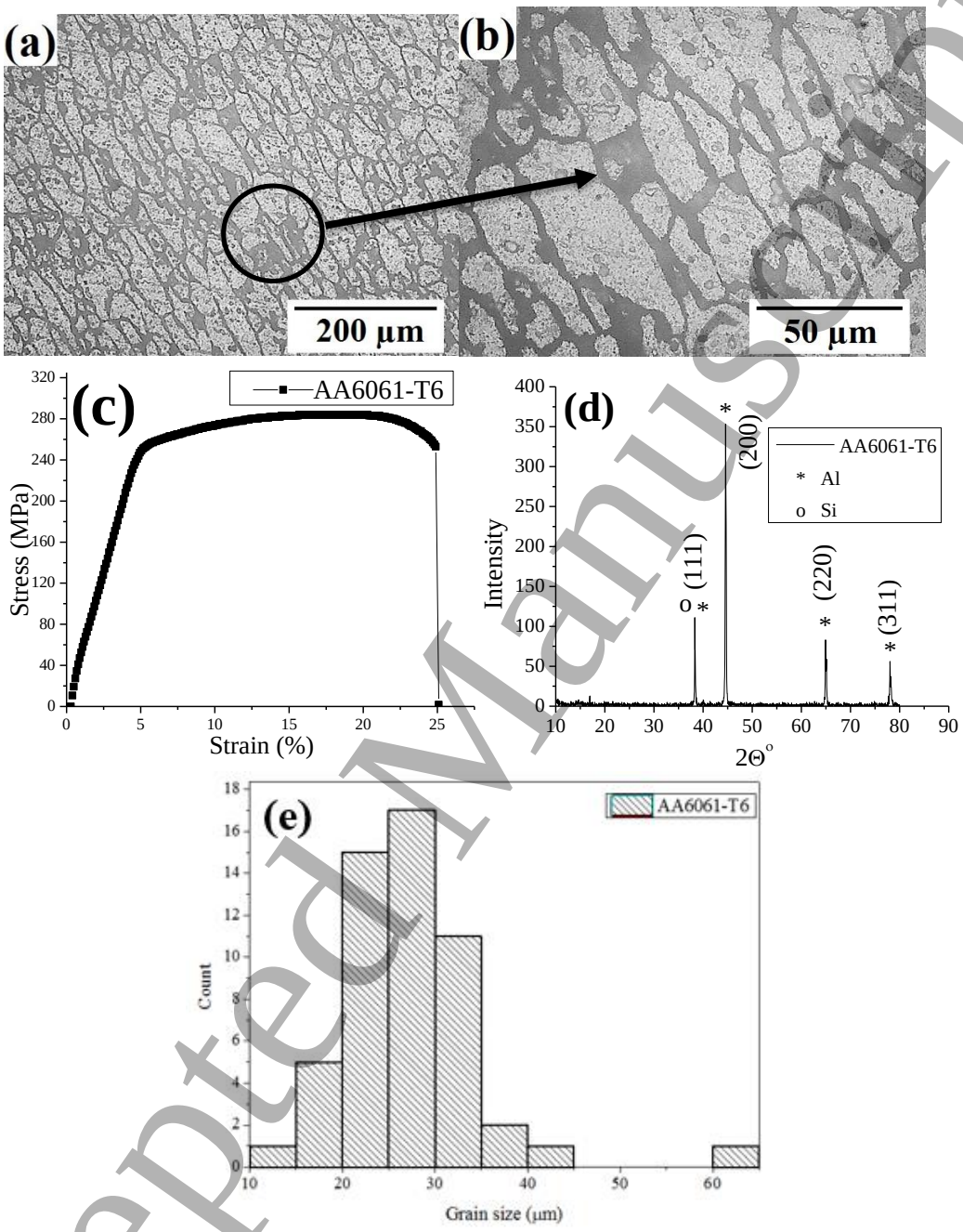


Figure 2. Substrate material (AA6061-T6) characterization (a & b) microscope image at 200 µm and 50 µm respectively (c) Stress vs strain plot (d) XRD plot (e) Grain size (µm)

CMT welding machine (TPS400i) by Fronius (as shown in figure 3) is used for performing the experiments, as per the design of experiment (DOE) given in table 2. Taguchi L9 design matrix is used for experiments, which helps in finding the optimal parameters while performing minimum number of trials, which are presented in table 3. It gives the best weld bead in terms of penetration, overall appearance of the bead, high strength with low HAZ, less amount of residual stresses, negligible deformation of the substrate material for thin sheets and high gap bridging ability. Current, voltage and WFR are

Accepted Manuscript

the dependent parameters in CMT and MIG P. Changing the value of one-parameter leads to change in the other as well due to synergic lines.

Table 3. Design matrix

S.No.	I (A)	S (mm/sec)
1.	80	7.5
2.	80	10.5
3.	80	13.5
4.	100	7.5
5.	100	10.5
6.	100	13.5
7.	120	7.5
8.	120	10.5
9.	120	13.5

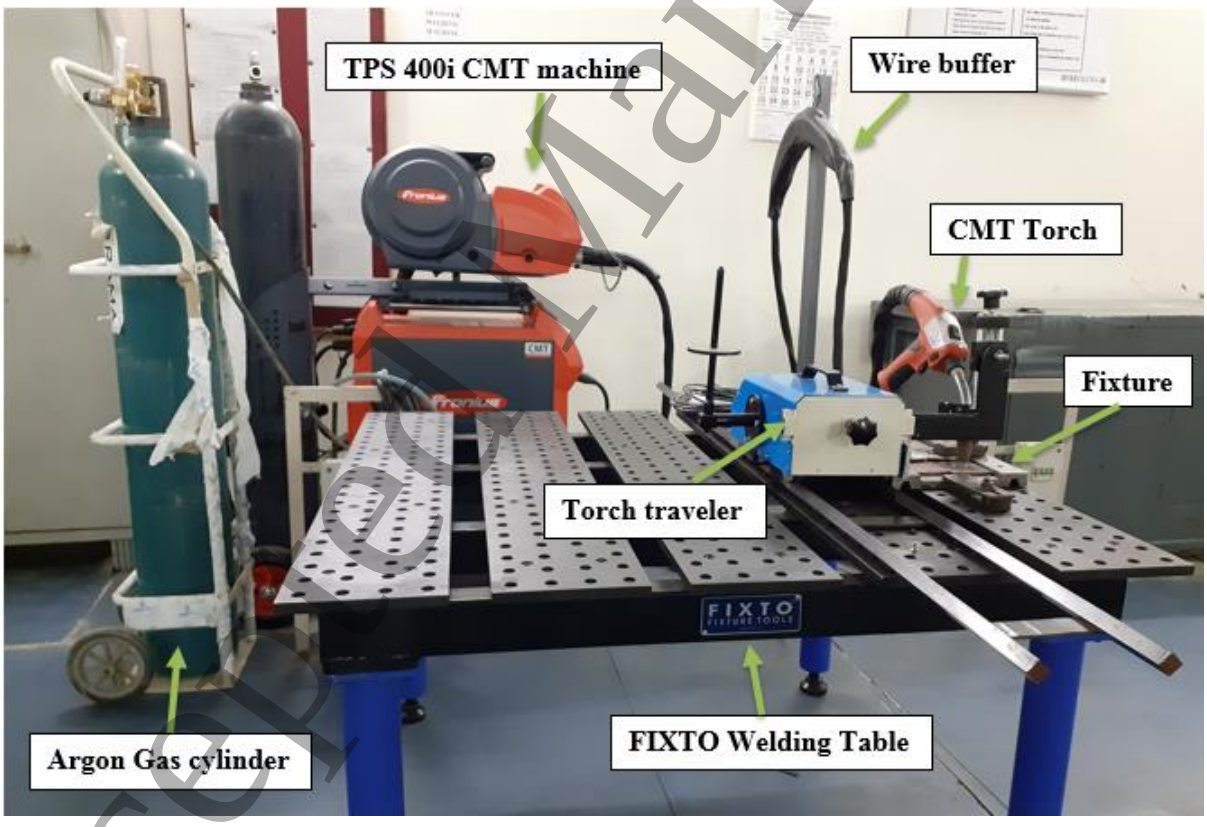


Figure 3. Experimental setup for CMT machine

Schematic diagram of fixture containing AA6061-T6 plate of dimension 100 mm x 60 mm x 3.18 mm ~~as~~ is shown in figure 4. Before welding, each plate is cleansed with acetone ($(CH_3)_2CO$). In order to eliminate the surface films and various other impurities, metal surface is subsequently cleaned with steel wire brush. After cleaning the surface, sample was fastened in

Accepted Manuscript

a fixture as displayed in figure 4, with torch angle placed at 90 º, CTWD is 10mm, stick-out is 5mm and argon (99.99 % purity) as a shielding gas at flow rate of 15 L/min are kept constant.

The standard bead-on-plate procedure was used to lay the weld beads on AA6061-T6 plates by using a wire of Ø1.2 mm of ER4043 (AlSi5 %). For every process parameters listed in table 2; five such beads were laid on the substrate material.

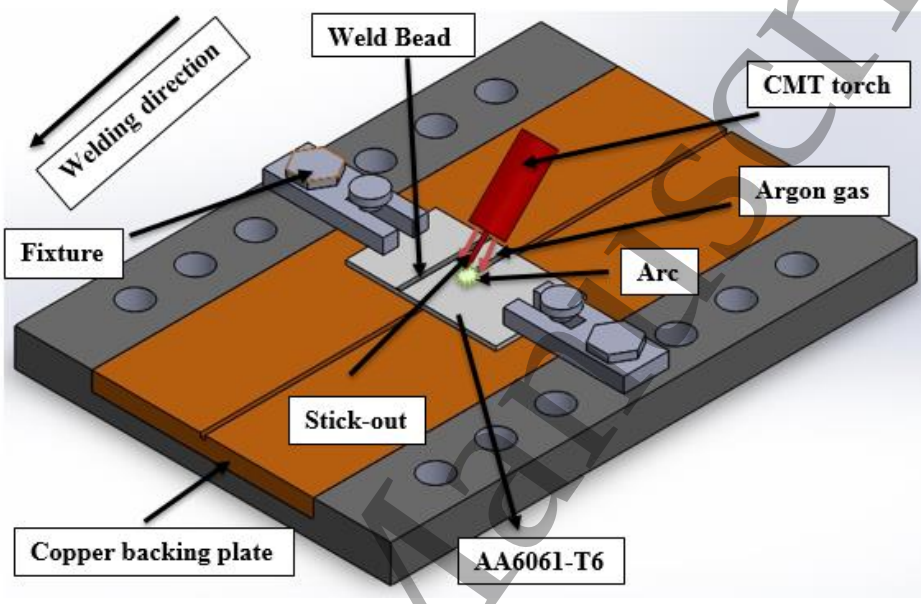


Figure 4. Schematic diagram of fixture containing AA6061-T6 plate of dimension 100 mm x 60 mm x 3.18 mm

After making the weld bead, each plate was cut from the middle parallel to the length. It was then dry polished with the use of different grades of emery paper (320, 400, 600, 800, 1000, 1200, 1500, 2000 and 2500) and etched using keller's reagent (1 ml HF, 1.5 ml HCl, 2.5 ml HNO_3 and 95 ml of water for 20 sec). The bead profiles were then captured at 10X magnification, bead dimensions and bead areas were measured by using a sipcon multi sensor CNC inspection system (SVI-5300-CNC-VT). Fisherscope (HM2000S) was used to measure the micro-hardness at various position (bead, BM and FZ). It has a micro indenter which indents with force 300 mN for 20 seconds having 5 seconds of creep time. Pulstec μ-X360n Full 2D High-Resolution X-ray Diffraction (HR-XRD) machine is used for measurement of residual stresses at various position of the weldment. Device specification contains a standard Cr X-ray tube (30 kV and 1 mA) with a collimator size of φ1 mm and measurement conditions. It is based on Cosα method, which acquire a full Debye-Scherrer ring (reveals grain orientation (texture) & grain coarsening, etc.); by a single short duration X-ray exposure from a 2-D detector [28]. It is a faster process because it does not require sample tilts at multiple angles as compared to traditional $\sin^2\psi$ technique. In this method, X-ray exposure is upto 1 μm into the material and measure the fundamental atomic plane spacing and difference in spacing as a result of processing. This measured lattice spacing is collected by Debye ring with a single measurement using a 2-D detector.

3. Results and Discussion

3.1. Weld bead dimensions

Weld bead dimensions give the overall view of the bead geometry. In weld bead geometry, reinforcement height (R), penetrated depth (P), weld width (W) and angle of contact (Θ) are measured and percentage dilution (D) is calculated with the help of sipcon multi sensor CNC inspection system. Weld bead dimensions is important to be measured to understand

Accepted Manuscript

the insight view of the bead geometry that helps in the overall cost of the welding. Good and economical weld joint needs deeper penetration for higher tensile strength; lower HAZ, weld width and weld reinforcement for lesser consumption of filler wire [15].

3.2.Macrostructure of weld bead in cross-sectional direction

Macrostructure in the cross-sectional direction of the weldment clearly shows the amount of pores that are present in the weld, depth of penetration, reinforcement, HAZ and BM. Table 4, 5 and 6 shows the macro-images of cross-section of CMT, MIG pulse synergic and MIG manual weld bead respectively. It can be seen that MIG P and MIG M is having high number of pores, which is formed due to gases entrapment when bead undergoes fusion. At such high heat inputs in MIG M and MIG P, aluminium alloys are susceptible to solubility of gases, majorly hydrogen which creates porosity. In contrast to heat input, alloying element is also responsible for hydrogen solubility. Mg element in aluminium alloys has high hydrogen affinity because the interactions between the Mg atoms and hydrogen atoms are stronger than those between aluminium and hydrogen [29]. Different welding zones are also visible after polishing and etching of the samples. Welding process parameters has variable effect on the weld bead, as current increases while keeping welding speed constant all the weld geometry increases due to high amount of high input. As welding speed increases while keeping current constant, penetration, weld width and dilution is decreased due to low heat input. Figure 5 (B & C), shows the crack formation on the plates of MIG P and MIG M. When heat input is 249.45 J/mm (120 A, 7.5 mm/sec) for MIG P and 271.64 J/mm (120 A, 7.5 mm/sec) & 194.74 (120 A, 10.5 mm/sec) for MIG M, cracks are seen on the surface of the bead with high amount of deposition. This elevated heat input and rapid cooling of the molten pool produces high amount of tensile residual stresses, which leads to formation of cracks on the surface of the bead. CMT is a welding process of low heat input is able to produce beads free from cracks with the same process parameters as shown in figure 5(A). It is noticed that the oxide layer deposition (aluminium oxide) i.e. white layer adjacent to the weld bead, is having more thickness in MIG P and MIG M as compared to CMT. Black soot, which is highly shown in MIG P and MIG M plates of 120 A, is magnesium oxide, which is formed by vaporized magnesium due to arc's heat.

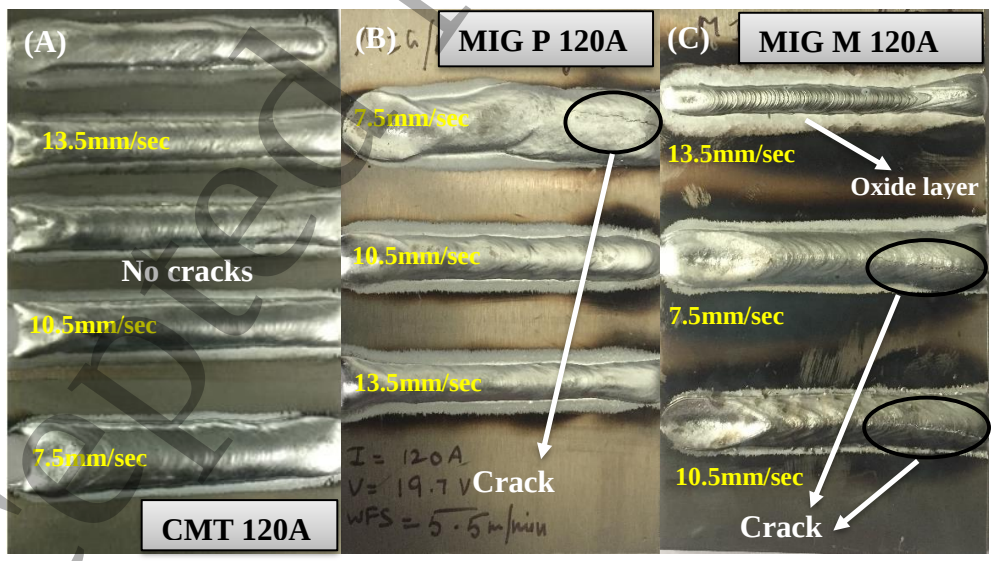


Figure 5. (A) No crack on CMT 120 A; Crack formation on bead on plate (B) MIG P (120 A) (C) MIG M (120 A)

Accepted Manuscript

Table 4. Macro-images of CMT weld bead

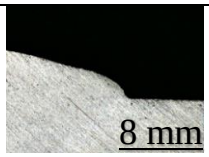
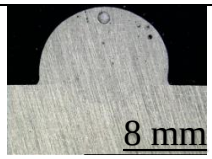
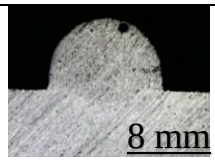
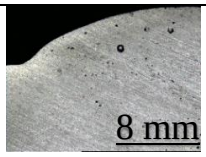
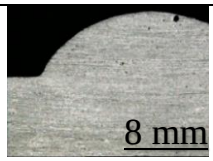
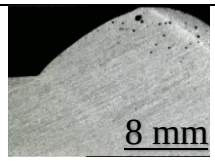
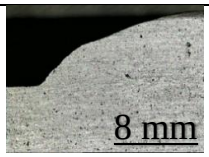
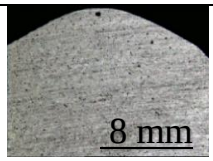
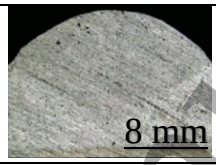
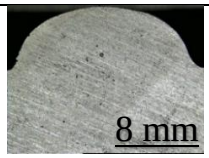
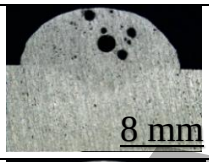
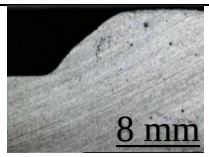
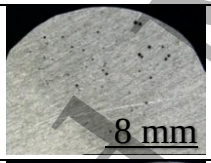
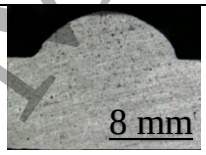
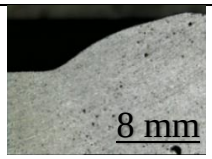
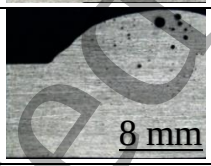
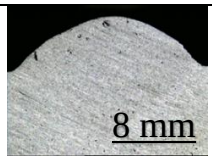
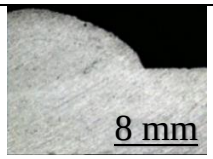
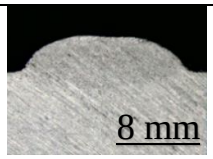
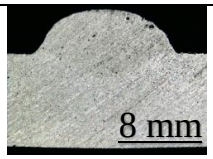
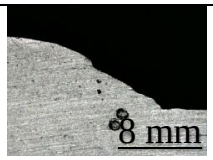
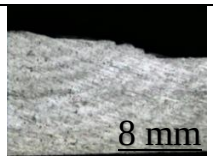
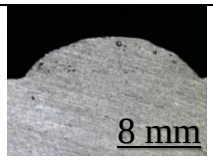
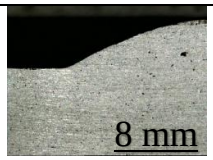
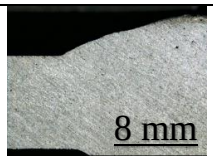
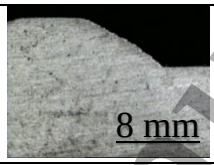
	7.5mm/sec	10.5mm/sec	13.5mm/sec
80A			
100A			
120A			

Table 5. Macro-images of MIG P weld bead

	7.5mm/sec	10.5mm/sec	13.5mm/sec
80A			
100A			
120A			

Accepted Manuscript

Table 6. Macro-images of MIG M weld bead

	7.5mm/sec	10.5mm/sec	13.5mm/sec
80A			
100A			
120A			

The samples were subjected to X-ray diffraction for the identification of the formation of intermetallic phases. As the parent material and filler material majorly constituting aluminum and silicon so primary peaks depicts the same as the major elements as shown in figure 6, which is confirmed by EDX plots. XRD plots shows lesser amount of intermetallic in CMT as compared to MIG P and MIG M owing to its low heat input. XRD plot of CMT depicts only Mg_2Si and Al_{12} Mg_{17} binary phases, which exists at lower temperature of 580 ^oC and 450 ^oC respectively. Fe_2Si binary phase is formed in both MIG P and MIG M, which exists at a temperature of about 1100 ^oC – 1200 ^oC due to their high thermal heat input. More amount of intermetallic phases causes lower mechanical properties which results in strength degradation.

Accepted Manuscript

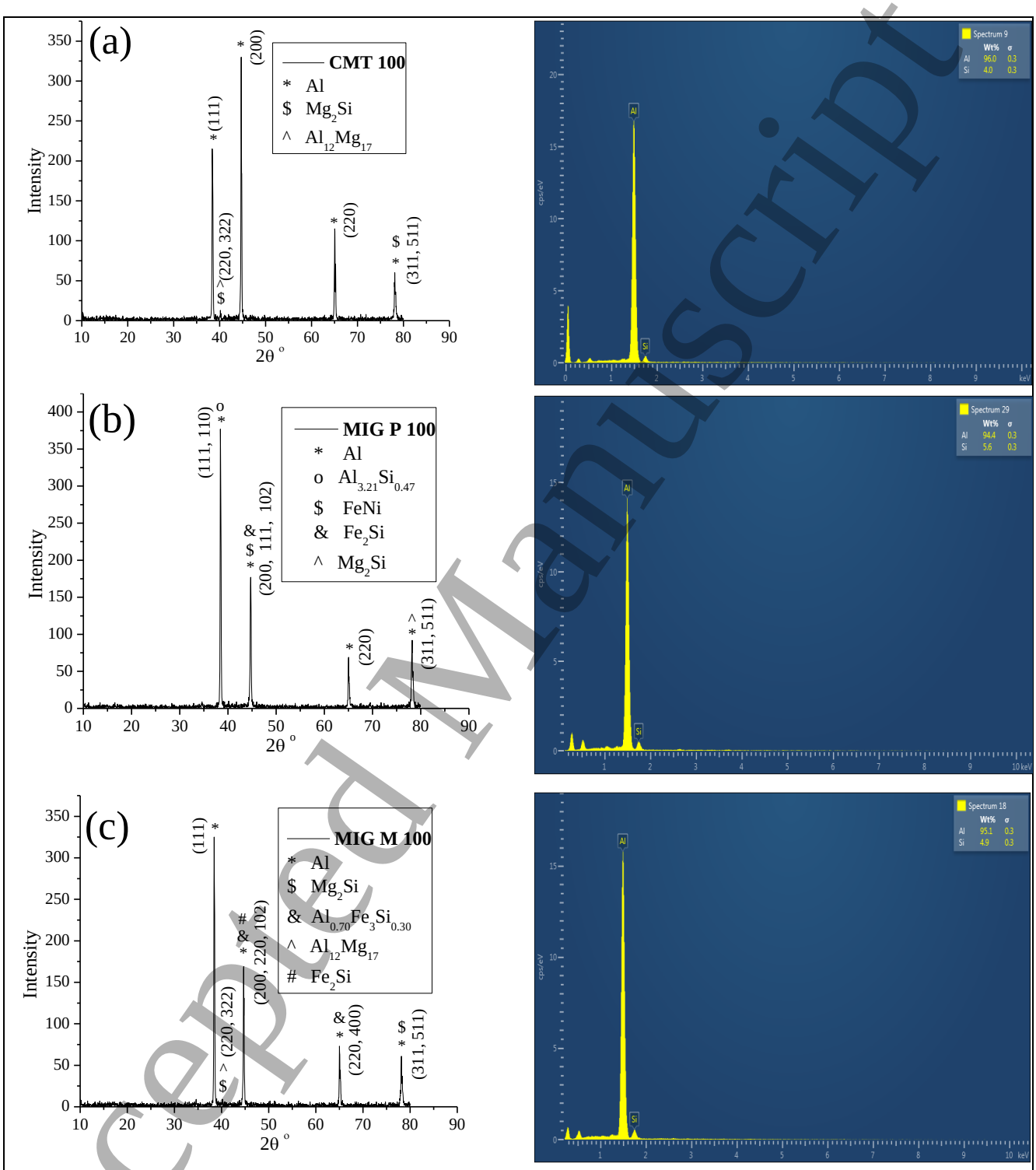


Figure 6. XRD and EDX plots for (a) CMT-100, (b) MIG P-100 and (c) MIG M-100.

Accepted Manuscript

3.3.Microstructure of weld bead

Olympus GX41 compact inverted metallurgical microscope is used for microstructural images. Figure (7), (8) and (9) shows the microstructural images from optical microscope for CMT, MIG P and MIG M respectively. Three samples were taken for examining the microstructure from each welding processes. Microstructure is taken at the (a) Base Metal (BM), (b & c) Fusion Zone (FZ) and (d) Weld Zone (WZ). Image (b) clearly shows the fusion line where AA6061-T6 substrate material and ER4043 filler metal fuses. It can be clearly noticed from Figure 7(b), 8(b) and 9(b) the difference in the microstructure of parent metal and weldment, which is separated by fusion line (shown in black dashes). Grain boundaries and grain structure is clearly shown in Figure 7(a), 8(a) and 9(a) which is representing parent metal. At higher current (i.e. 120 A) values it is noticed that high amount of black spot is seen in the parent metal and bead region, which is due to high heat input. It causes non-uniformity in cooling rate, which results in high amount of brittle compounds (Mg_2Si) that is seen as the black spots. Mg_2Si dissolves in FZ of the weld bead and results in large precipitation and coarsening of grains [30, 31]. These black spots deteriorates the mechanical properties and surface profile of the bead. It is more in MIG M process as compared to CMT and MIG P. High amount of pores are present in MIG P (figure 8) and MIG M (figure 9) as compared to CMT, which weakens the joint, and results in reduced tensile strength. CMT is showing better microstructures as compared to MIG M and MIG P owing to its spatter free welds even at high current inputs and better bead aesthetics. FESEM help us examine the morphology and microstructure. Figure 10 shows the FESEM images of fusion zone of CMT at 100 A current with 10.5 mm/sec of welding speed. Figure 10(b) shows large precipitates of dissolved Mg_2Si in FZ.

Accepted Manuscript

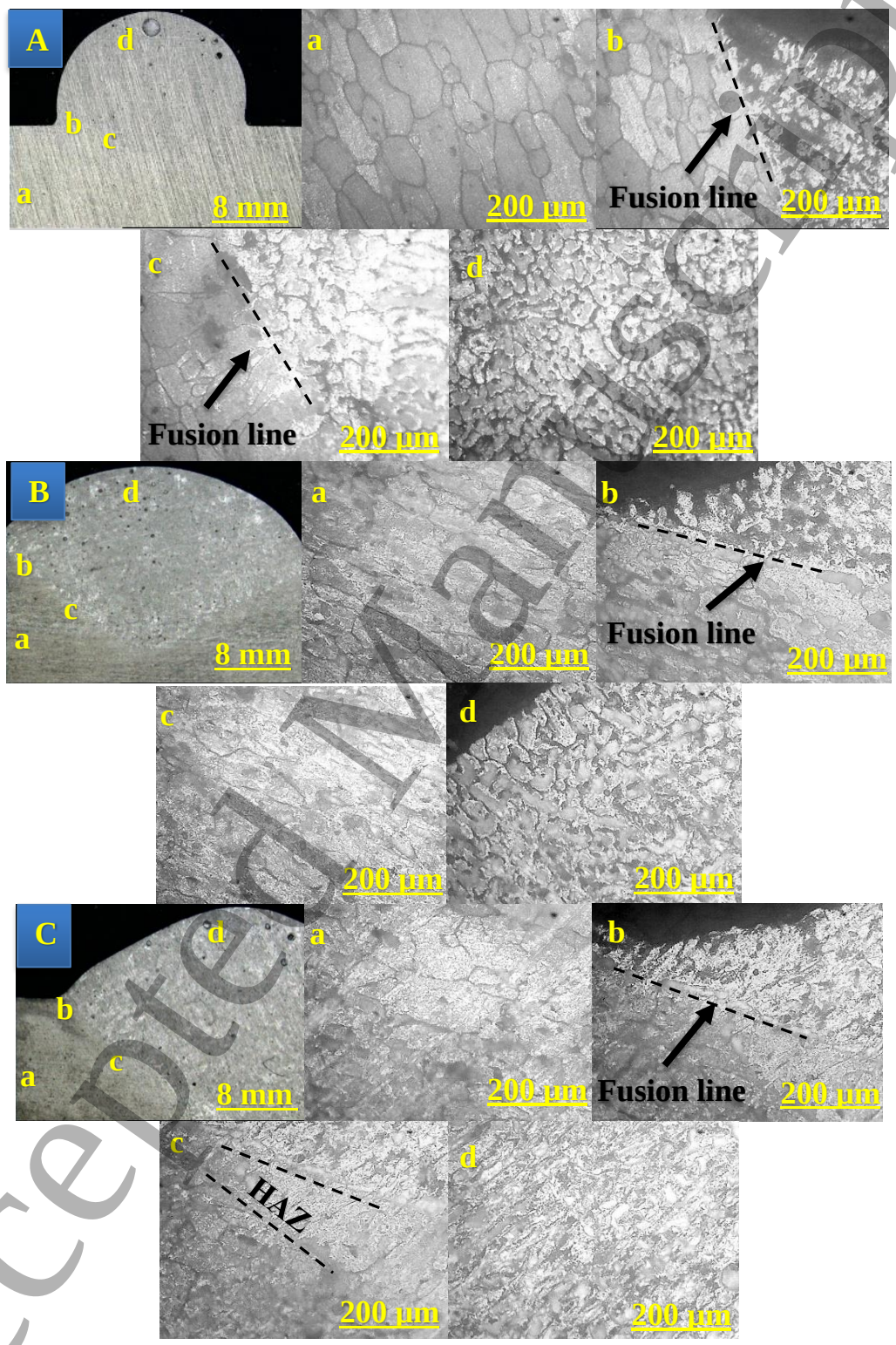


Figure 7. Vision inspection image A,B and C of CMT 80 A, 100 A and 120 A respectively at 10.5 mm/sec. Optical microscope image at 500 X (200 µm) (a) BM (b & c) FZ (d) WZ

Accepted Manuscript

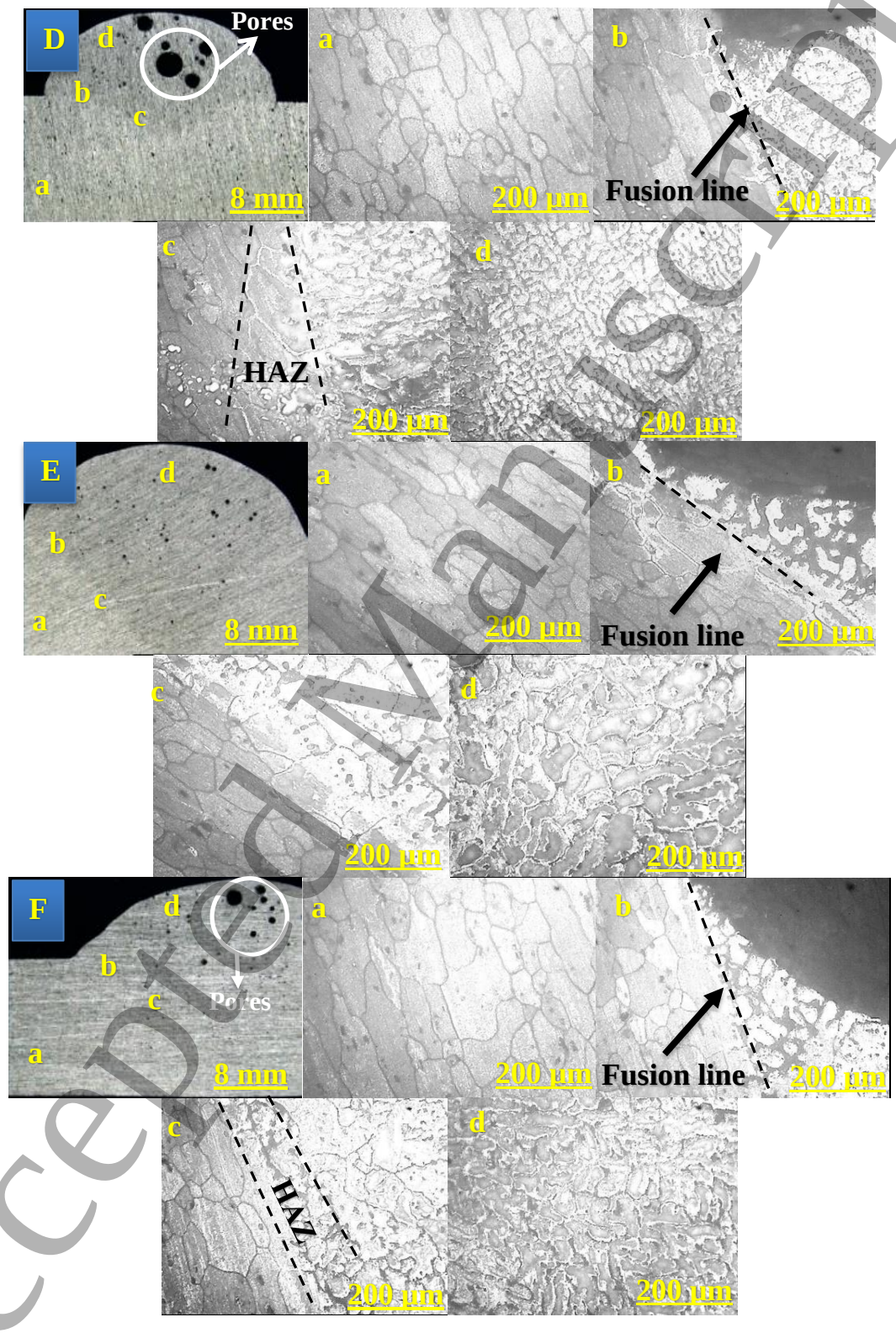


Figure 8. Vision inspection image D, E and F of MIG P for 80 A, 100 A and 120 A respectively at 10.5 mm/sec. Optical microscope image at 500 X (200 µm) (a) BM (b & c) FZ (d) WZ

Accepted Manuscript

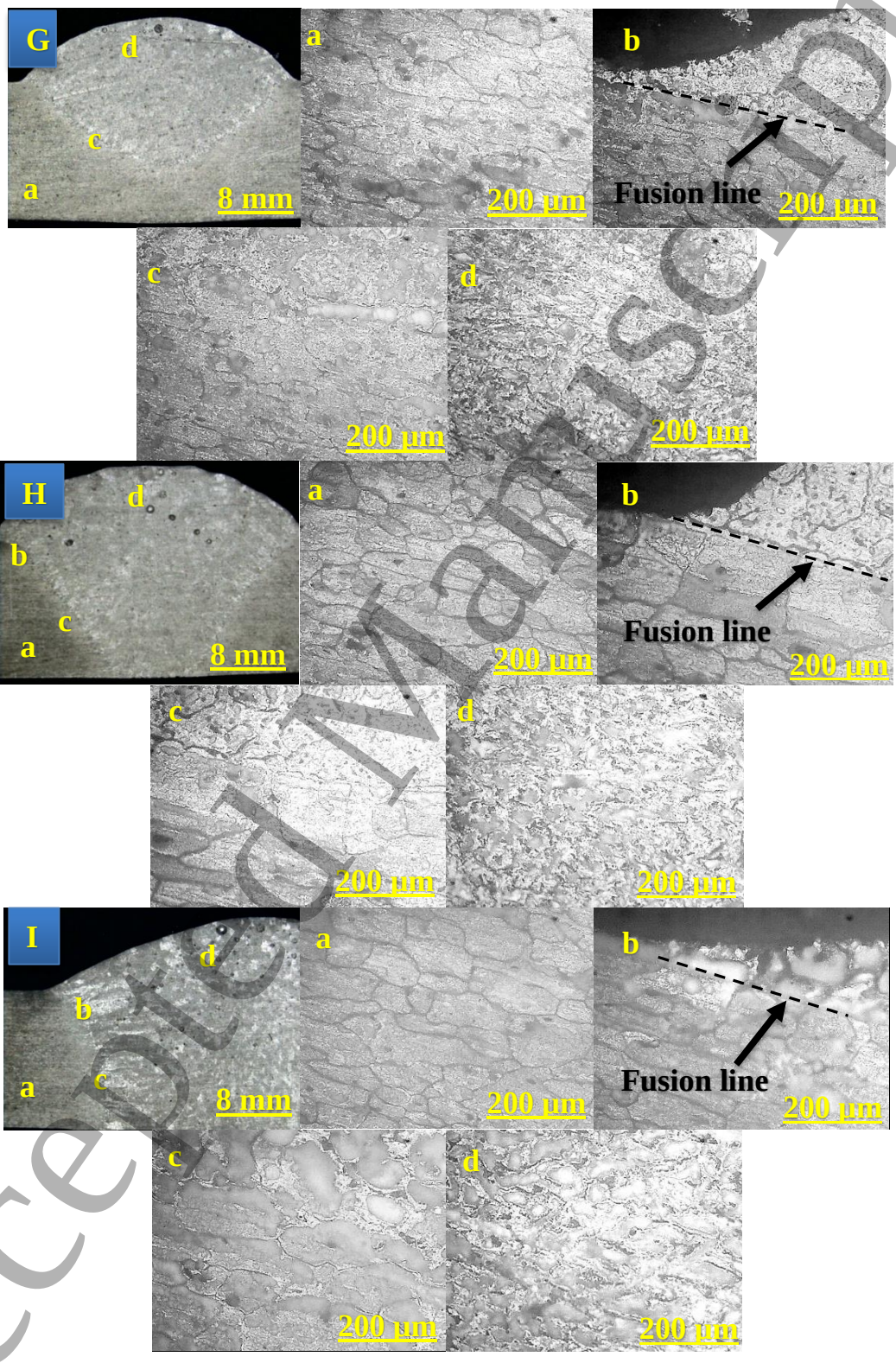


Figure 9. Vision inspection image G, H and I of MIG M for 80 A, 100 A and 120 A respectively at 10.5 mm/sec. Optical microscope image at 500 X (200 µm) (a) BM (b & c) FZ (d) WZ

Accepted Manuscript

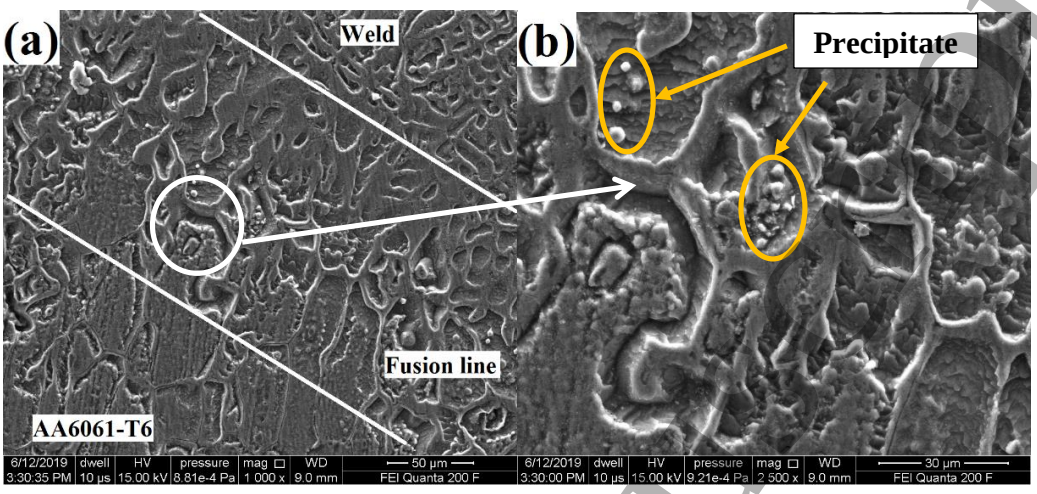


Figure 10. FESEM images of fusion zone of CMT 100 A at (a) 50 µm scale and (b) 30 µm scale

FESEM clearly shows the difference among the grain structure that is formed due to non-uniformity in temperature. This is majorly seen in the weld bead and fusion zone where high heat input is experienced.

3.4. Effect on Dilution (D) and Heat Input (H)

Dilution (D) and heat input (H) are directly proportional to each other, as heat input on the substrate material increases, more volume being melted, causing increase in dilution and vice versa. As mentioned above, dilution is the fraction of penetration area (B) to the area of total weld metal (A+B) as shown in equation (1). It is the factor, which is influenced by primary dimensions of the weld bead. Figure 11 shows the heat input curve and figure 12 shows the dilution curve between CMT, MIG P and MIG M respectively w.r.t. welding speed (S). Heat input is calculated with the help of equation (4). It is an important response in terms of influencing various factors like plate deformation, thickness of HAZ, arc stability, etc.

Accepted Manuscript

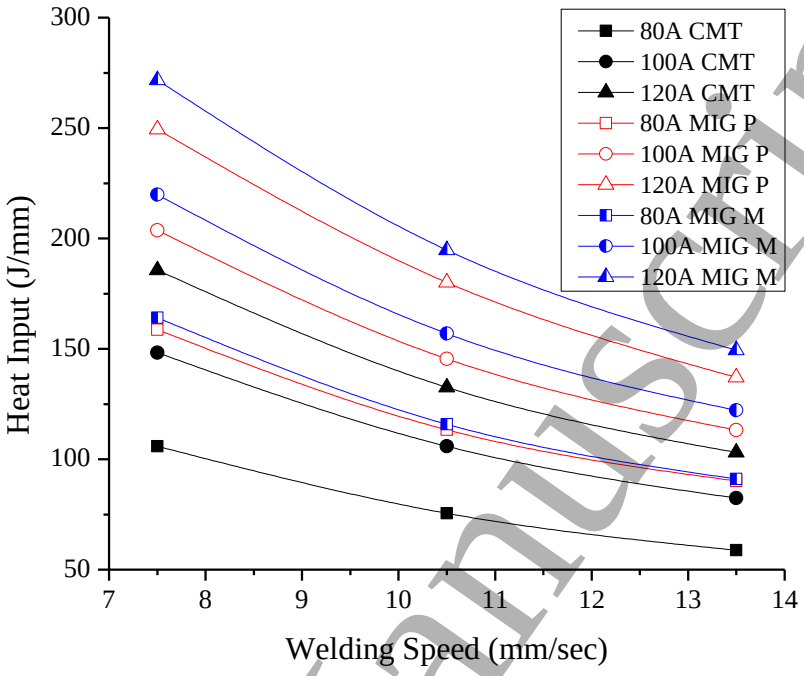


Figure 11. Comparison of heat input (J/mm) between CMT, MIG P and MIG M

With increase in welding speed causes less dilution, which can be attributed to the lower heat input with increased weld speed. From figure 12, dilution (D) experiences a decreasing trend with increase in the value of welding speed (S) for all the processes, because as S increases 7.5-13.5 mm/sec, heat input on the weld decreased as shown in figure 11, and less amount of substrate material melted thereby decreasing the penetration area as shown by macro-images. Similarly, by lowering down the value of S, heat input on the substrate material increases and more amount of substrate material melted which cause deeper penetration resulting in high percentage of dilution. As current increases from 80 A to 120 A, Dilution increases for all the welding processes. Higher current produces high heat input (as shown in figure 11) causes melting of substrate material for deeper penetration resulting in increment in percentage of dilution. It can be seen from figure 12, as current increases from 80 A to 100 A, D increases to about 40 % for CMT, 50 % for MIG P and 60 % for MIG M in comparison to 10 % for CMT, 12 % for MIG P and 16 % for MIG M when current is increased from 100 A to 120 A w.r.t welding speed. So, for a current period of 80-100 A, D is showing higher values as compared with current period of 100-120 A because heat input for current period 80-100 A is more as compared with 100-120 A as shown in figure 11. Figure 13, clearly depicts that the trend for dilution w.r.t heat input. This graph shows, as the current increases from 80-100A for all the processes, there is significant change in the dilution due to variable change in heat input. Comparatively, CMT is having less amount of dilution and heat input on various process parameters from MIG P and MIG M. It is a much stable process in terms of dilution and heat input, as it is not showing any variable change in the output. Lower amount of dilution and heat input is needed for fabricating a good joint, which is having high strength by minimizing the HAZ.

Accepted Manuscript

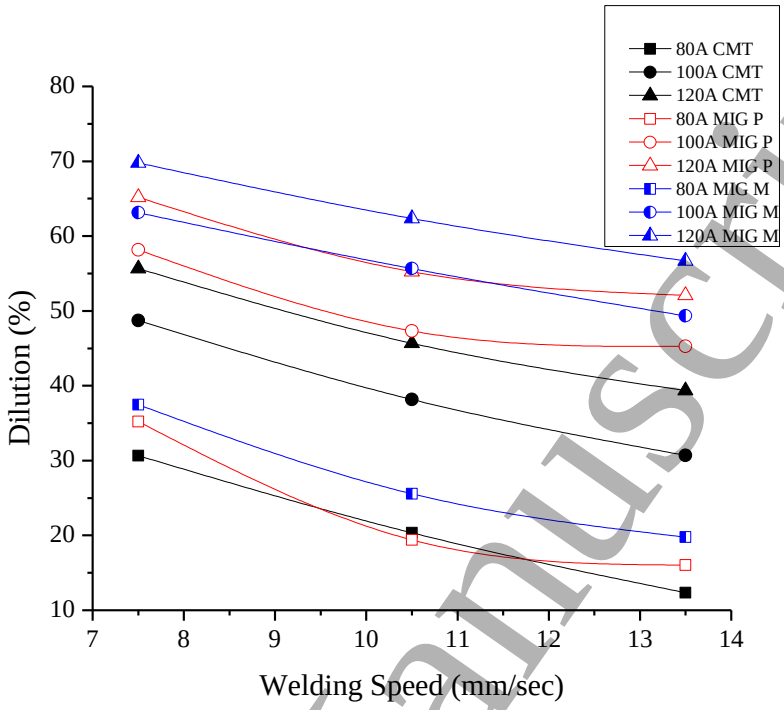


Figure 12. Comparison of dilution (%) between CMT, MIG P and MIG M

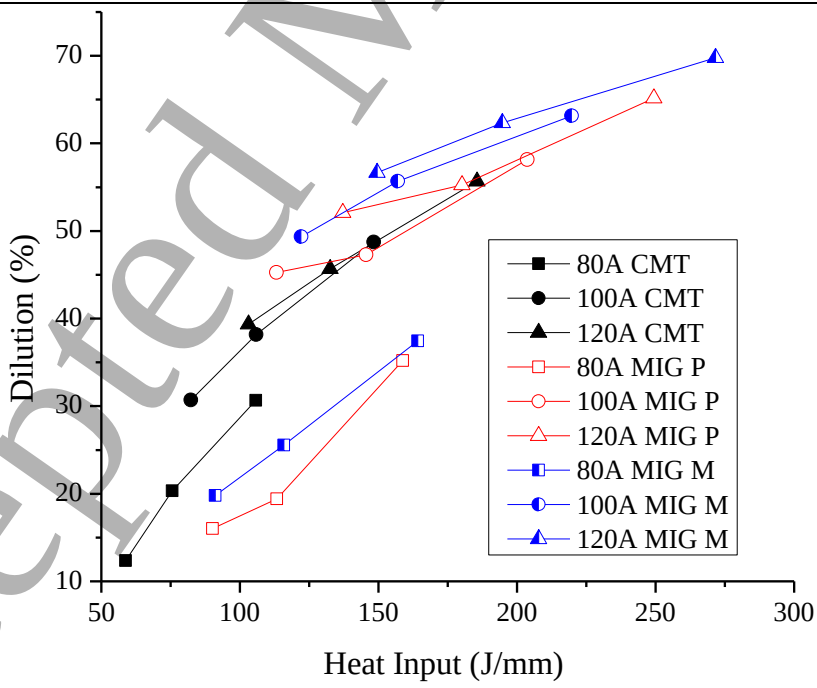


Figure 13. Dilution (%) vs heat input (J/mm) between CMT, MIG P and MIG M

Accepted Manuscript

3.5. Effect on Penetration

Penetration is an important factor for having a good joint efficiency, deeper penetration results in higher joint efficiency due to filler material fusing into the substrate material. Deeper penetration with higher tensile strength is required for a good quality joint provided its HAZ, weld reinforcement and width are lower to reduce weld metal consumption are vital requirement for all types of welding [15]. It is usually noted that, as is obvious in figure 14, penetration is positively influenced by an increase in current and a decrease in welding speed. CMT process exhibits steeper decrement of 60 %, 45 % and 35 % in 80 A, 100 A and 120 A respectively as welding speed increases from 7.5 to 10.5 mm/sec. In comparison to MIG P and MIG M, it is 40 %, 16 % and 14 % in 80 A, 100 A and 120 A respectively. MIG P and MIG M shows deeper penetration of about 45 % in comparison to only 20 % in CMT when current increases from 80 to 100 A for speed of 7.5 mm/sec owing to high amount of heat input at higher current and lower welding speed. Wire feed rate (WFR) too plays an imperative role in increasing the penetration. With increase in WFR, current increases, which helps in raising the heat input for a weld duration, leading in high amount of melting of substrate material. In CMT, the oscillatory movement of the wire helps in controlling the penetration by maintaining the arc length. With higher WFR in CMT, the current increases which helps in more droplet detachments, results in higher penetration. It is observed from figure 15, CMT is achieving similar amount of penetration as compared to MIG P and MIG M with lesser amount of heat input, which ultimately results in saving of energy.

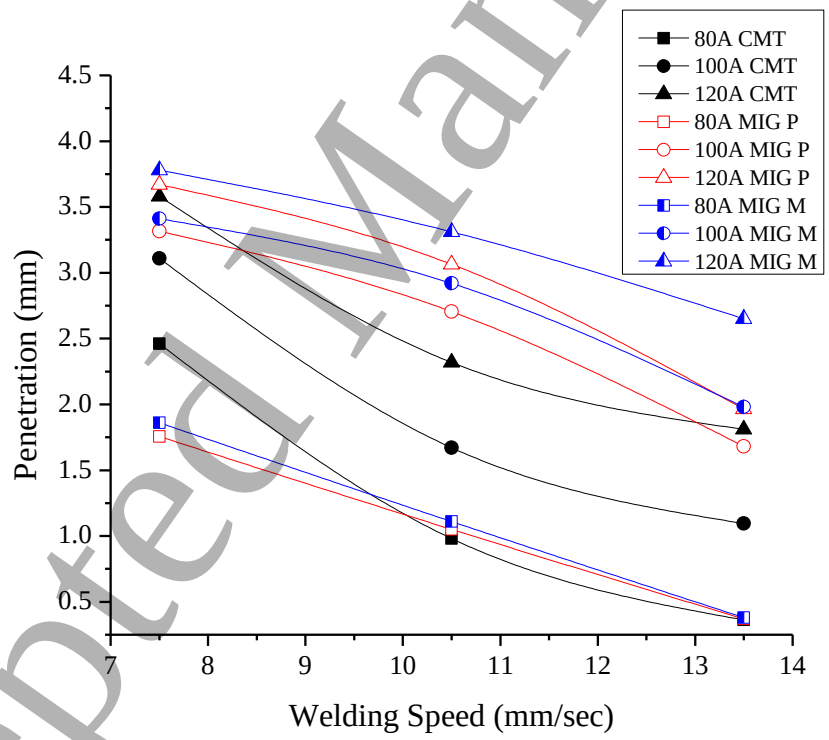


Figure 14. Comparison of penetration (mm) between CMT, MIG P and MIG M

Accepted Manuscript

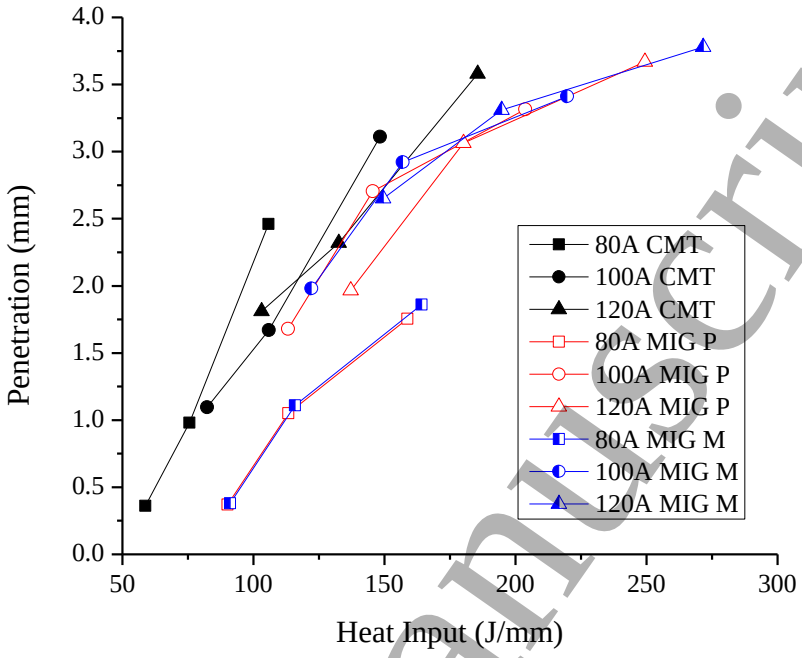


Figure 15. Penetration (mm) vs heat input (J/mm) between CMT, MIG P and MIG M

3.6. Effect of Residual Stress

Residual stress magnitude depends on several factors such as deposited weld bead size, weld sequence, total deposited weld metal volume, weld geometry, deposited weld metal and adjacent BM strength, and cooling rate. As shown in the figure 16, CMT shows less residual stress compared to MIG P and MIG M. Due to high heat input and cooling rate, FZ has higher compressive stress for all the fusion welding process (CMT, MIG P and MIG M) compared to the weld bead. CMT is showing 6-12 % and 21-29 % decrement at the beads compared with MIG P and MIG M respectively. At FZ, CMT is experiencing, 11-14 % and 17-25 % decrement as compared with MIG P and MIG M respectively [32]. Weld bead and fusion zone has compressive stress, which has a positive impact on the tensile and fatigue strength. The top and bottom surfaces of the weld joint experiences a greater cooling rate during welding than the middle part of the weld and HAZ. This creates differential expansion and contraction by welding in the plate's thickness. Metal contraction near the surface begins even if the middle portion of the material is still in solidus form. This contributes to the growth of the residual compressive stress at the top surface and the residual tensile stress at the middle of the weld bead. Therefore, compressive residual stresses are deliberately caused to improve the fatigue behavior of mechanical parts, while attempts are made to decrease residual tensile stresses using multiple methods such as post-weld heat treatment, shot peening, spot heating, etc. As tensile residual stress, causes crack nucleation and further crack propagation under tensile load condition, which results in tensile failure etc. As the current level increases and the welding speed decreases, residual stress increases, resulting in the crack propagation and base plate distortion.

Accepted Manuscript

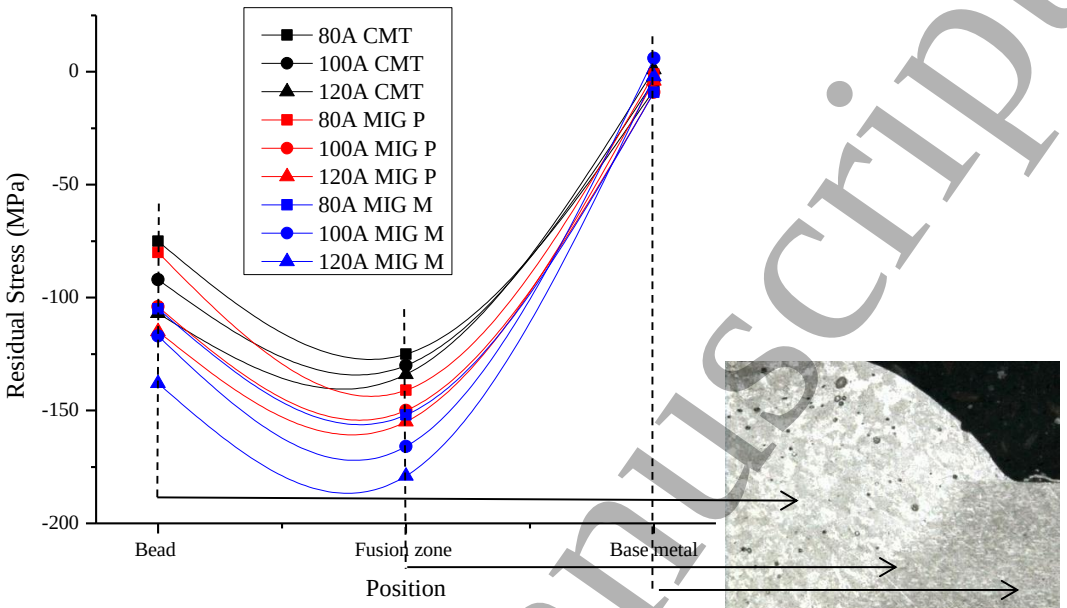


Figure 16. Residual stresses induced on different position of the specimen

Figure 17 shows the residual stress profile with respect to alpha angle for CMT, MIG P and MIG M welding process at different positions for 100 A and 10.5 mm/sec process parameters. The residual stress peaks for all the samples are in between 155 º to 160 º. Red shade peaks are seen in fusion zone and weld bead, which indicated higher concentration of residual stresses. Blue shade peaks means negligible amount of residual stresses. From these residual profile peaks, FWHM is measured at different position of the weld bead by the machine. Thicker the residual peaks, higher will be the value of FWHM that results in finer grain structure (i.e. higher micro-hardness) as shown is figure 18. It is seen that, FWHM and micro-hardness are linearly related and directly proportional to each other. The welded samples were segmented and polished with various grades of waterproof emery paper (320, 400, 600, 800, 1000, 1200, 1500, 2000 and 2500) and etched using standard keller's reagent to identify FL and FZ. A diamond micron indenter determines and presents the hardness values and evaluates them with the help of a microscope. The indenter indents for 20 sec, 5 sec for creep and 20 sec for returning the indenter. As shown in figure 18, weld bead is having low hardness w.r.t BM and FZ. Compared to MIG P and MIG M, CMT has higher hardness values. It shows a 17.5 % increase in FZ compared to BM. CMT shows an increase in FZ hardness of 5 % compared to MIG P and MIG M. Microhardness at weld bead for MIG welding is in the range of 45-65 HV [33]. Obtained microhardness values for CMT are higher than MIG welding. The variation in micro-hardness only depends on the zone of the weld bead. The material generally loses its original strength during solidification in the fusion zone (FZ) owing to the strain hardening effect. Formation of brittle intermetallic compounds (IMC) in the FZ stimulates an increase in micro-hardness in the case of aluminum alloys.

Accepted Manuscript

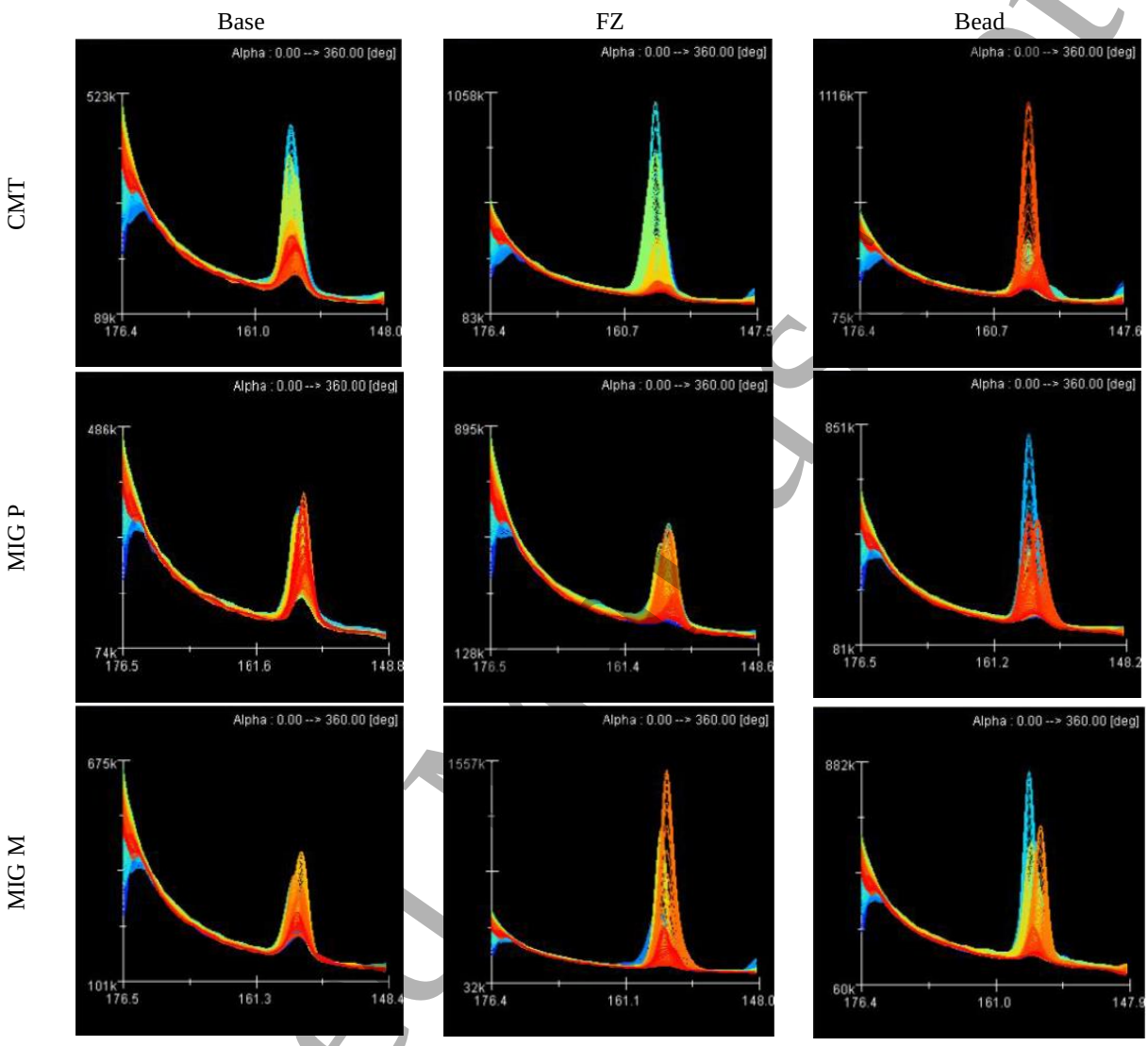


Figure 17. Residual stress profile vs alpha angle for CMT, MIG P and MIG M at different positions for 100A and 10.5mm/sec process parameters

Accepted Manuscript

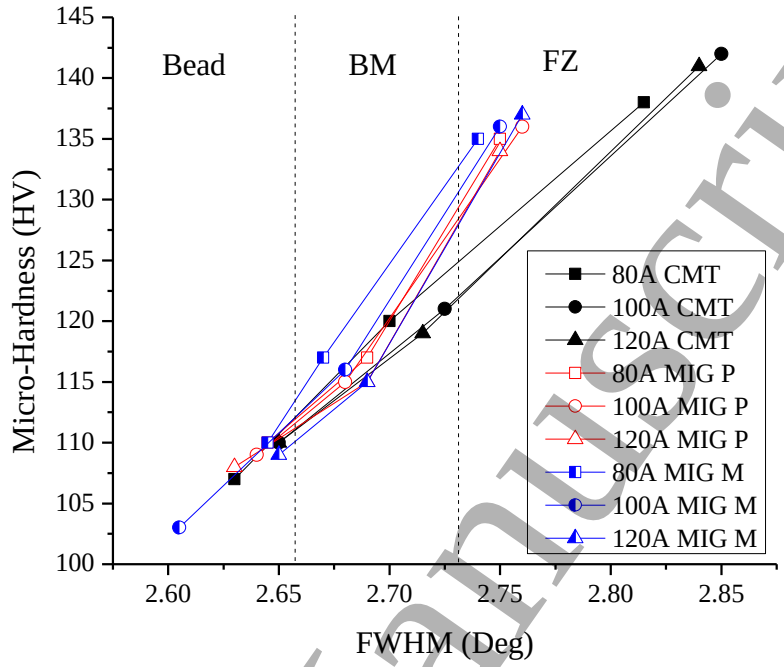


Figure 18. Microhardness vs FWHM

Accepted Manuscript

4. Conclusions

1. Overall, CMT has relatively lower penetration for the same current and welding speeds compared to MIG pulse synergic and MIG manual, which helps to weld thin sheets. A good amount of penetration and dilution with low heat input is required for a better joint efficiency. As current increases from 80 A to 100 A, Dilution increases to about 40 % for CMT, 50 % for MIG Pulse synergic and 60 % for MIG Manual in comparison to 10 % for CMT, 12 % for MIG P and 16 % for MIG M when current increases from 100 A to 120 A w.r.t welding speed due to high input.
2. High heat input and rapid cooling in the molten pool produce high residual stresses that are seen on the bead's surface in the form of cracks, which ultimately reduces the joint strength. These cracks are experienced by MIG pulse synergic and MIG manual, which CMT prevents due to low heat input. As a result, compared to MIG Pulse and MIG Manual, CMT has low residual stresses.
3. At the beads, CMT shows a decrease in residual stress of 6-12 % and 21-29 % compared to MIG Pulse synergic and MIG Manual respectively and in the fusion zone, CMT is experiencing a decrease of 11-14 % and 17-25 % compared to MIG Pulse synergic and MIG Manual.
4. When FWHM values increase, grain size decreases, resulting in higher hardness values. Compared to MIG P and MIG M, CMT has higher hardness values resulting in finer grain structure. It shows a 17.5 % increase in FZ compared to BM. CMT shows a 5 % increase in FZ hardness compared to MIG P and MIG M.

Accepted Manuscript

References

1. Elrefaey A and Ross NG 2015 Microstructure and mechanical properties of cold metal transfer welding similar and dissimilar aluminum alloys *Acta Metall. Sin-Engl.* 28 715-724.
2. Mcauley JW 2003 Global sustainability and key needs in future automotive design *Environ. Sci. Technol.* 37 5414-5416.
3. Selvamani ST, Govindarajan P, Ajaymohan M, Hariharan SJ, and Vigneshwar M 2018 Correlation between Micro hardness and Microstructure of CMT Welded AA 7075 Al Alloy *In IOP Conference Series: Materials Science and Engineering* 390 012058.
4. Totten GE and MacKenzie DS 2003 Handbook of aluminum: vol. 1: physical metallurgy and processes. *CRC press*, New York.
5. Kaufman JG 2000 Introduction to aluminum alloys and tempers *ASM international*, USA.
6. Mathers G 2002 The welding of aluminum and its alloys *Woodhead publishing*, England.
7. Mossman MM and Lippold JC 2002 Weldability testing of dissimilar combinations of 5000-and 6000-series aluminum alloys *Weld. J.* New York 81 188-194.
8. Kumar NP, Chinnadurai T, Vendan SA, and Shanmugam NS 2018 Techno Economical Evaluation for Energy Analysis in AA6061 during Cold Metal Transfer Welding *Materials Today: Proceedings* 5(11) 23375-23383.
9. Kumar NP, Vendan SA, and Shanmugam NS 2016 Investigations on the parametric effects of cold metal transfer process on the microstructural aspects in AA6061 *J. Alloy Compd.* 658 255-264.
10. Selvi S, Vishvaksenan A, and Rajasekar E 2018 Cold metal transfer (CMT) technology-An overview *Def. Technol.* 14 28-44.
11. Zhang HT, Feng JC, He P, Zhang BB, Chen JM and Wang L 2009 The arc characteristics and metal transfer behaviour of cold metal transfer and its use in joining aluminum to zinc-coated steel *J. Materials Science & Engineering* A499 111–3
12. Evangeline A and Sathiya P 2019 Cold metal arc transfer (CMT) metal deposition of Inconel 625 superalloy on 316L austenitic stainless steel: microstructural evaluation, corrosion and wear resistance properties *Mater. Res. Express* 6 066516.
13. İrizalp AO, Durmuş H, Yüksel N and Türkmen İ 2016 Cold metal transfer welding of AA1050 aluminum thin sheets *Matéria (Rio de Janeiro)* 21 615-622.
14. Kolahan F and Heidari M 2010 A new approach for predicting and optimizing weld bead geometry in GMAW *International Journal of Mechanical Systems science and engineering* 2 138-142.
15. Esme U, Bayramoglu M, Kazancoglu Y and Ozgun S 2009 Optimization of weld bead geometry in TIG welding process using grey relation analysis and Taguchi method *Materiali in tehnologije* 43(3) 143-149.
16. Choudhary A, Kumar M, & Unune DR 2019 Experimental investigation and optimization of weld bead characteristics during submerged arc welding of AISI 1023 steel *Def. Technol.* 15 72-82.
17. Hunt AC, Kluken AO, and Edwards GR 1994 Heat input and dilution effects in microalloyed steel weld metals. *Weld. J.* New York 73 9-s.
18. Sun YL, Obasi G, Hamelin CJ, Vasileiou AN, Flint TF, Balakrishnan J, Smith MC, Francis JA 2019 Effects of dilution on alloy content and microstructure in multi-pass steel welds *J. Mater. Process. Tech.* 265 71-86.
19. Saha MK, Hazra R, Mondal A and Das S 2019 Effect of Heat Input on Geometry of Austenitic Stainless Steel Weld Bead on Low Carbon Steel *Journal of The Institution of Engineers (India): Series C* 100(4) 607-615.
20. DuPont JN and Marder AR 1996 Dilution in single pass arc welds *Metall. Mater. Trans. B* 27 481-489.
21. Sankar BV, Lawrence ID, and Jayabal S 2018 Experimental Study and Analysis of Weld Parameters by GRA on MIG Welding *Materials Today: Proceedings* 5 14309-14316.
22. Sakthivel R, Venkadeshwaran P, Sridevi R, Meeran RA, and Chandrasekaran K 2016 Effect of Welding Current, Arc Voltage and Gas Flow Rate on Depth of Penetration during MIG Welding of AA2014 Plate *way* 2 30.
23. Utkarsh S, Neel P, Mahajan MT, Jignesh P, and Prajapati RB 2014 Experimental investigation of MIG welding for ST-37 using design of experiment *International Journal of Scientific and Research Publications* 4 1.

Accepted Manuscript

24. Khanna P, and Maheshwari S 2018 Development of mathematical models for prediction and control of weld bead dimensions in MIG welding of stainless steel 409M *Materials Today: Proceedings* 5 4475-4488.
25. Sharma SK, Maheshwari S, and Singh RKR 2019 Effect of heat-input and cooling-time on bead characteristics in SAW *Mater. Manuf. Process* 34 208-215.
26. Galantucci LM, Tricarico L, and Spina R 2000 A quality evaluation method for laser welding of Al alloys through neural networks *CIRP Annals* 49 (1) 131-134.
27. Mendez PF and Eagar TW 2001 Welding processes for aeronautics *Advanced materials and processes* 159 39-43.
28. Kumar S, Grover S, and Walia RS 2018 Effect of hybrid wire EDM conditions on generation of residual stresses in machining of HCHCr D2 tool steel under ultrasonic vibration *International Journal on Interactive Design and Manufacturing (IJIDeM)* 12 1119-1137.
29. Anyalebechi PN 1995 Analysis of the effects of alloying elements on hydrogen solubility in liquid aluminum alloys *Scripta metallurgica et materialia* 33(8) 1209-1216.
30. Ahmad R and Bakar MA 2011 Effect of a post-weld heat treatment on the mechanical and microstructure properties of AA6061 joints welded by the gas metal arc welding cold metal transfer method *Materials & Design* 32(10) 5120-5126.
31. Maisonnette D, Suery M, Nelias D, Chaudet P, Epicier T 2011 Effects of heat treatments on the microstructure and mechanical properties of a 6061 aluminium alloy *Materials Science and Engineering: A* 528(6) 2718-2724.
32. Pinto H, Pyzalla A, Hackl H, and Bruckner J 2006 A comparative study of microstructure and residual stresses of CMT-, MIG-and laser-hybrid welds *In Materials science forum* 524 627-632.
33. Ishak M, Noordin NFM, Razali ASK, Hakim L, Shah A 2015 The effect of filler ER4043 and ER5356 on weld metal structure of 6061 aluminium alloy by Metal Inert Gas (MIG) *methanol yield over Bi2S3/CdS photocatalyst:4*.

Accepted Manuscript

International Journal of Engineering Trends and Technology (IJETT) - Volume 67 Issue 4 - April 2019

Knowledge Discovery and Data Mining Review Papers

Zenawi Haileslassie Berhe (MTech) *[1], Associate professor Vinod Kumar *[2]
*Department of Computer Engineering, Delhi Technological University, Delhi, India *[1]*
*Department of Computer Engineering, Delhi Technological University, Delhi, India *[2]*

Abstract

The main objective of this paper is on challenges in data mining and knowledge discovery and different types of knowledge discovery in database techniques like visualization, statistical, different types of machine learning, and different types of databases are highly used in knowledge discovery and data mining. Knowledge discovery and data mining also mainly focus on datamining technology, data mining tasks, different types of knowledge discovery in database or KDD process are the main thigs focused on this paper [1][2][3][4][5][6].

Keywords - *data mining tasks, knowledge discovery, Large Databases, KDD process*

I. INTRODUCTION

Knowledge discovery and data mining is most of the time focused on combinations of different necessary data points. And knowledge discovery and data mining satisfy the new things that is generated in our daily life of different computational techniques and different toolsused in different information's during extracting and assisting of different data mining points. These kinds of information's are also known as or considered as knowledge. On this paper mainly focused on how these different types of techniques are used and how the different types of tools and subject of the knowledge discovery in dataare knowledge discovery and data mining are used in processing different data mining actives and also different sets how to generate and how to extract with the help of data mining work [1].

In data mining and knowledge discovery there are different types of approaches and these different types of approaches can be categorized in to four main categories base on their operations of data mining techniques that are included in different techniques used in knowledge discovery and data mining as follows: -

The first approach is predictive model creation is one the supervised machine learning techniques used in different analyzing and also discover different producers' step by step in database.

The second type of approach in data mining and knowledge discovery is segmentation. Segmentation is also used in different types of techniques most of the time.

The third type of approaches is also detection deviation this is used most of the time use in different statistical techniques [1].

II. THE MAIN DIFFERENCE BETWEEN DATA MINING AND KNOWLEDGE DISCOVERY IN DATABASE AND ITS BASIC CONCEPTS

The main difference between data mining and knowledge discovery in database is data mining is most of the time used to analyze different types of executive summary and different types of important information's how to analyze and how to identify with different categories of data mining and knowledge discovery and their angles of the different techniques and their relationships [3].

Basically, knowledge discovery and data mining are also finding with different correlations and regressions data is extracted from huge database.

A. The main tasks of data mining are

The main tasks of data miningare focused with different kinds of techniques used and generated data from data mining and knowledge discovery in database (KDD) or huge amount of data.

And another basic task of data mining is also by classifying theses different types of techniques and tools how to organize them to generate or extract data from large amount of data in the data base [5].

1. Summarization

Summarization is the deliberation or speculation of information. A set of assignment pertinent information is condensed also, disconnected. This outcomes in a littler set which gives a general outline of the information, for the most part with total data.

2. Classification

Classification gathers a capacity or model which decides the class of an object dependent on its

attributes. A lot of articles are given as the preparation set. In it, each item is spoken to by a vector of traits alongside its class. a grouping capacity or model is developed by breaking down the connection between the characteristics and the classes of the items in the preparation set. This capacity or model would then be able to order future items. This causes us build up a superior comprehension of the classes of the protests in the database.

3. Clustering

Bunching recognizes classes additionally called bunches or gatherings for a lot of objects whose classes are obscure. The objects are clustered to the point that the interclass similitudes are augmented and the interclass similitudes are limited. This is done based on certain criteria characterized on the properties of the articles. When the bunches are chosen, the objects are marked with their relating bunches. The regular highlights for articles in a bunch are condensed to shape the class depiction.

4. Trend investigation

Time arrangement information are records aggregated after some time. For instance, an organization's deals, a client's Mastercard exchanges and stock costs are untouched arrangement information. Such information can be seen as articles with a property time. The items are previews of substances with qualities that change after some time. Finding the examples and regularities in the information advancements along the component of time can be entrancing.

B. Some of the steps of knowledge discovery in data base process

The knowledge discovery in data base process is the most type of interactive methods in different persuaders.

Knowledge discovery in data base (KDD)is divided in to different steps as follows: -

The first type of knowledge discovery in database (KDD) process is data selection and the second of knowledge discovery in database (KDD) process is data cleaning.

The third type of knowledge discovery in database (KDD) process is data transformation and the fourth type of knowledge discovery in database (KDD) process is pattern searching (data mining), and the last type of knowledge discovery in database (KDD) process is Finding presentation and also finding interpretation and finding evaluation[2][4].

Knowledge (Information) Discovery in Databases is the non-insignificant procedure of recognizing legitimate, novel, possibly helpful, and at last reasonable examples in information. Here information is a lot of actualities (e.g., cases in a database) and example is an articulation in a few languages portraying a subset of the information or a demonstrate pertinent to that subset. Removing a design assigns fitting a model to information, discovering structure from information, or all in all any abnormal state depiction of a lot of information. The term process infers that KDD is involved numerous means, which include information readiness, scan for examples, information assessment, and refinement, all rehashed in numerous emphases.

The found examples ought to be substantial on new information with some level of sureness. We likewise need examples to be novel (in any event to the framework, and ideally to the client) and conceivably helpful. At long last, the examples ought to be justifiable, in the event that not promptly, at that point after some post preparing.

The KDD procedure is intelligent and iterative, including various strides with numerous choices being made by the client. Here we extensively diagram a portion of its fundamental advances:

- ✓ Learning the application space:
  - Relevant earlier information and objectives of application
- ✓ Creating an objective informational index: information choice
- ✓ Data cleaning and preprocessing
- ✓ Data decrease and projection:
  - Find helpful highlights, dimensionality/variable decrease and invariant portrayal.
- ✓ Choosing the mining algorithm(s)
- ✓ Data mining: scan for examples of intrigue
- ✓ Interpretation: investigation of results.
  - Visualization, change, evacuating repetitive designs.
- ✓ Use of found learning are some of knowledge discovery procedures.

C. Some of Data mining challenges are

The most commonly challenges that are most of the time happen in different researches and also in the domains of knowledge discovery in data base is most of the time happened in our daily activities[1][3].

Basically, the most commonly challenges happened in data mining are the following: -

The first and basic and commonly happened challenge is normalization.

The second challenge in data mining is estimation missing value and the third challenge is detection outlier.

III. DATA MINING APPLICATIONS

The main applications of data mining are in short and précised form is the following things: -

Some of the applications of data mining is advertising, different bioinformatics, in different Customer Relationship Management (CRM), database Marketing, E-Commerce, health Care, investment/Securities, different manufacturing, Process Control, different sports activities and Entertainment, telecommunications, different web etc.… [6].

Information mining systems have been connected effectively in numerous territories from business to science to help the following additional things: -

A. Business applications

Many associations presently utilize information mining as a mystery weapon to keep or pick up a focused edge. Information mining has been utilized in database promoting, retail information investigation, stock determination, credit endorsement, and so forth.

B. Science applications

Data mining systems have been utilized in cosmology, sub-atomic science, prescription, geography and some more.

C. Other applications

Data mining systems have likewise been utilized in medicinal services the executives, charge extortion location, and illegal tax avoidance observing and even games.

IV. CONCLUSION

This paper is focused on data mining and knowledge discovery in database (KDD) and what are the challenge, applications and what are the tasks and process of knowledge discovery and data mining in short and also what are the different types of techniques in data mining, knowledge discovery and its basic concepts[4][5].

REFERENCES

[1] Mykhaylo Lobur1, Yuri Stekh2, Vitalij Artsibasov3" challenges in knowledge discovery and data mining in datasets",IEEE Expert, Vol. 9, No. 1,1994, pp. 60–66.

[2] Qi Luo1, 2 "advancing knowledge discovery and data mining", 0-7695-3090-7/08 $25.00 © 2008 IEEE

[3] M.Lobur, Yu. Stekh, A.Kernytskyy, Faisal M.E. Sardieh" Some Trends in Knowledge Discovery and Data Mining"MEMSTECH'2008, May 21-24, 2008, Polyana, UKRAINE.

[4] M-Tahar Kechadi1, and Ilias K. Savvas2 1University College Dublin, Ireland, 2T.E.I. of Larissa, Greece, tahar.kechadi@ucd.ie, savvas@teilar.gr" Cooperative Knowledge Discovery & Data Mining CKDD" 2012 IEEE 21st International WETICE.

[5] M.Lobur, Yu. Stekh, A.Kernytskyy, Faisal M.E. Sardieh" Some Trends in Knowledge Discovery and Data Mining"MEMSTECH'2008, May 21-24, 2008, Polyana, UKRAINE.

[6] Manish Marwah1, Ratnesh Sharma1, Naren Ramakrishnan2, Cullen Bash1, Chandrakant Patel1 1HP Labs, Palo Alto, CA 2Virginia Tech, Blacksburg, VA"Knowledge Discovery and Data Mining for Enhanced Sustainability of Physical Ecosystems"ITHERM'08, June 2008, Orlando, FL.

Land Use Land Cover Classification from Satellite Imagery using mUnet: A Modified Unet Architecture

Lakshya Garg[1], Parul Shukla[2], Sandeep Kumar Singh[2], Vaishangi Bajpai[2] and Utkarsh Yadav[2]
[1]*Electrical and Electronics Engineering, Delhi Technological University (DTU), Delhi, India*
[2]*Strategic Operations and Research, RMSI Pvt Ltd Noida, Delhi, India*

Keywords: Satellite Imagery, Land-use-classification, Convolutional Networks, Remote Sensing, Deep Learning.

Abstract: Land-use-land-cover classification(LULC) is used to automate the process of providing labels, describing the physical land type to represent how a land area is being used. Many sectors such as telecom, utility, hydrology etc need land use and land cover information from remote sensing images. This information provides an insight into the type of geographical distribution of a region with providing low level features such as amount of vegetation, building area, and geometry etc as well as higher level concepts such as land use classes. This information is particularly useful for resource-starved rapidly developing cities for urban planning and resource management. LULC also provides historical changes in land-use patterns over a period of time. In this paper, we analyze patterns of land use in urban and rural neighborhoods using high resolution satellite imagery, utilizing a state of the art deep convolutional neural network. The proposed LULC network, termed as **mUnet** is based on an encoder-decoder convolutional architecture for pixel-level semantic segmentation. We test our approach on 3 band, FCC satellite imagery covering 225 km^2 area of Karachi. Experimental results show the superiority of our proposed network architecture vis-à-vis other state of the art networks.

SCIENCE AND TECHNOLOGY PUBLICATIONS

1 INTRODUCTION

Recent advancements in remote sensing have resulted in easy accessibility of satellite imagery of Earth. Among many applications of satellite imagery, Land-use-land-cover (LULC) forms an integral part. LULC plays a pivotal role in urban planning, land resource management and provides a useful insight into the growth rate indexes of different population spectrum. LULC models also highlight the historical changes in an area by quantitatively showcasing the changes in habitation, vegetation, water areas over a span of years. Recent breakthroughs in deep learning have resulted in LULC's application to various new versatile domains such as development of smart cities, urban planning, environmental monitoring and disaster recovery. The remote sensing community benefited the most with the use of deep convolutional neural networks for various tasks such as automatic feature extraction for classifying road, building footprints, grasslands etc (Mnih, 2013). Land use classification refers to the consolidation of physical land attributes defining a region and what the cultural and socio-economic function that the land serves. In this paper, we address LULC classification by modifying the Unet (Ronneberger et al., 2015) architecture for pixel level segmentation. The proposed model is henceforth referred to as **mUnet**. mUnet consists of a ladder-like structure consisting of convolutional encoder layers followed by a series of decoding convolutional layers. Our method has the following advantages as compared to the traditional segmentation methods.

- mUnet provides classification at pixel level.
- mUnet consists of less number of trainable parameters as compared to original Unet.
- mUnet outperforms other state of the art segmentation architectures.

Our paper consist of the following sections: In section-II we review the studies which adopt deep learning methods to the problem of land-use-classification. In section-III, we present the dataset

Garg, L., Shukla, P., Singh, S., Bajpai, V. and Yadav, U.
Land Use Land Cover Classification from Satellite Imagery using mUnet: A Modified Unet Architecture.
DOI: 10.5220/0007370603590365
In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2019)*, pages 359-365
ISBN: 978-989-758-354-4
Copyright © 2019 by SCITEPRESS – Science and Technology Publications, Lda. All rights reserved

curated. In Section-IV, we examine the proposed methodology including network architecture and training details. In section-V we showcase our results and findings henceforth followed by a conclusion in section-VI.

2 RELATED WORK

The literature on land-use-land-cover classification can be broadly classified into two categories: patch-based and pixel-based methods. Patch-based approach (Papadomanolaki et al., 2016) (Albert et al., 2017) (Tong et al., 2018) is based on predicting labels for an image-patch whereas pixel-based approach (Kussul et al., 2017) (Torres et al., 2017) the labels are predicted for each underlying pixel of the image. Patch-based approaches are primarily based on neural-networks such as VGG (Deng et al., 2009), ResNet (He et al., 2016), AlexNet (Krizhevsky et al., 2012) for classifying the image. Pixel-based approaches on the other hand use Deep convolutional Neural nets combined with many up-sampling layers in order to accomplish semantic segmentation of images in an end to end manner.

2.1 Patch-based-Approach

Papadomanolaki et al (Papadomanolaki et al., 2016) evaluated deep learning frameworks based on convolutional neural networks for the accurate classification of spectral remote sensing data. A patch-based approach was utilized with a patch size of 28 28 pixels, consisting of 4 and 6 band images. The publicly available data of SAT-4 and SAT-6 images were utilized to fine tune the three different models Alexnet (Krizhevsky et al., 2012), Alexnet-small(Iandola et al., 2016), VGG(Deng et al., 2009). SAT-4 images were classified into four different classes (barren land, trees, grassland, water bodies) while SAT-6 consisted of 6 different classes (barren land, trees, grasslands, road, buildings and water bodies). Similarly Albert et al (Albert et al., 2017) utilized a pre-trained Resnet (He et al., 2016) and VGG-16 (Deng et al., 2009) to classify patches into 10 predefined classes (water, airport, sports and leisure, low, medium, high density urban, green urban, forest, agriculture). Google maps imagery was utilized to train the networks and the data were taken across six European countries. A patch size of 224 224 3, 1.2m/px was given as input to the model. Xin-Yi et al (Tong et al., 2018) utilized a Patch-based CNN classification using a re-trained Resnet-50 network. The network had a total of 49 convolution layers, consisting of 16 bottleneck structures. 50 Gaofen-2 imagery of 1m/px resolution were used to train the network. The original images were divided into patches of 256 256 and 512 512 patches respectively and were given as input to the model.
It may be noted that misclassification is high in case of patch-based approach. Patch-based approach classifies an entire patch as a single label which may be incorrect if the patch consists of pixels belonging to other classes as well

2.2 Pixel-based-Approach

Pixel based approach mitigates the limitations of patch-based approach by predicting labels for each pixel in an image, thus reducing the misclassification rate. Kussul et al (Kussul et al., 2017) used unsupervised neural network techniques on hyper-spectral imagery. The authors used 1D and 2D CNN ensemble. 1D CNN ensemble looked across the pixel values along the spectral domain, while the 2D network architecture performed the spatial convolutions. Dataset used was a multi temporal, multi-Source data of the Kyiv region acquired from Landsat-8 and Sentinel 1-A. Eleven pre-distinguished classes were defined to classify the temporal bounds of the image. The authors use an unsupervised learning approach which inherently requires a large amount of training data. Torres et al (Torres et al., 2017) utilized a supervised learning approach, which involves training a fully convolutional neural (FCN) network in order to perform LULC classification at pixel-level on Dar Es Salaam dataset. The Images were broken down into 440 440 pixel tiles and then classified into 10 pre-defined classes. The authors used a FCN network which consist of an encoder architecture followed by a final up-sampling layer succeeded by a classifier, however it has been found that an encoder followed by a decoder which consists of trainable convolutional layers performs better in segmentation tasks (Ronneberger et al., 2015). Additionally, (Torres et al., 2017) (Kussul et al., 2017) used low resolution satellite imagery (30m/px, 10m/px) for LULC classification. In this paper we propose an encoder-decoder architecture for LULC classification on a high resolution satellite imagery (2m/px).

3 DATASET USED

Data used for training the network is Resourcesat 2 Satellite imagery of Karachi covering 225 km^2. LISS IV merged with Cartosat Satellite imagery was

used in the study because of its finer spatial resolution than other commonly used images such as Multi-Spectral Scanner (MSS). Figure 5 shows the full-scene FCC (false color composite) satellite imagery (7506 columns, 7475 rows). The 3 band imagery was acquired from National Remote Sensing Agency, Department of Space, Government of India. The image quality was rather good with relatively no cloud cover over the study area. The images have a spatial resolution of 2 m and 8 bit pixel depth. In this paper four classes are considered for Land use land cover classification, namely Habitation, Vegetation, Open and water. The training data is prepared for 26 km^2 and prediction is done over a complete study area covering 225 km^2 of Karachi. The image preprocessing and post processing is done using ArcGIS software.

4 METHODOLOGY

In the following section we detail our proposed approach.

4.1 Image Pre-processing

The ground truth thematic image consisting of four land use land cover classes is used as ancillary data for creating labels and validating as seen in figure 6. Vectors are generated corresponding to each class by raster to polygon conversion. These vectors are further used for preparing training datasets for 26 km^2 area of Karachi. The false color composite (FCC) satellite imagery is tilled to 36 images of 850 850 pixels wide and corresponding WKT (well known text) file is generated by superimposing vectors over the tiles figure 4 shows the training data labels as binary images generated from WKT file. The WKT file contains polygons co-ordinate information corresponding to each of the four classes and for each of the 36 tiles used for training.

4.2 Network Architecture

We propose a modified encoder-decoder network based on the Unet model which was was originally utilized for bio-medical segmentation (Ronneberger et al., 2015). The structure resembles, as the name signifies, alphabet U. Unet concatenates the encoder feature maps to up-sampled feature maps from the decoder at every stage to form a ladder-like structure. Our network (mUnet) starts with a 5-block convolutional base of neural units alternated with a max pooling layer. Each convolutional layer is followed by a ReLU activation function. The obtained feature maps are fed into a decoder. The decoder up samples the feature maps to produce predictions at the original spatial resolution. The decoder starts with a 5 block convolutional base of neural units with each convolutional layer preceded by an up-sampling layer. Each convolutional layer is followed by a ReLU activation function. Thus the feature maps finally obtained are fed into a fully 2D convolutional layer consisting of 4 classes with sigmoid activation function as shown in the figure 1.

Our network (mUnet) is inherently different from the traditional (Unet) in the following ways.

- mUNet has 19 convolutional layers as compared to UNet which has 23 convolutional layers.
- Filter size of 5 5 is used in mUNet while the traditional UNet has a filter size 3x3.
- The total number of trainable parameters in mUnet are less as compared to the Unet.
- The convolution operations used in mUnet are padded convolutions while in Unet there is spatial reduction between subsequent convolutional layers.

These features make our model much more suitable for LULC classification. For training the model a TITAN-XP GPU with CUDA Core 3840 was used. We tested our model on 32 Core, 2.10 GHz CPU.

4.3 Training Details

For training the model, cross-entropy loss function is used. Given a class c, the cross-entropy loss is computed as:

$$L = -(y_{ijc} \ln(p_{ijc}) + (1 - y_{ijc}) \ln(1 - p_{ijc})) \quad (1)$$

where the total number of classes are C (in our case 4). mUnet generates predicted binary map for each class. c. p_{ijc} represents the predicted activation value for pixel $I(i, j)$ and class c. For a given image I, of height H and width W, the ground truth consists of a tensor y of size H x W x C such that $y_{ijc} = 1$ if pixel $I(i, j)$ belongs to class c, otherwise $y_{ijc} = 0$. The network is trained using Adam optimizer (Kingma and Ba, 2014) for 200 epochs. A batch size of 64 is used along with a learning rate of 0.001.

4.4 Post Processing

The model returns binary png and pgw (geo-reference information) files corresponding to four classes (habitation, water, vegetation and open). Each binary file is then converted into corresponding four feature class

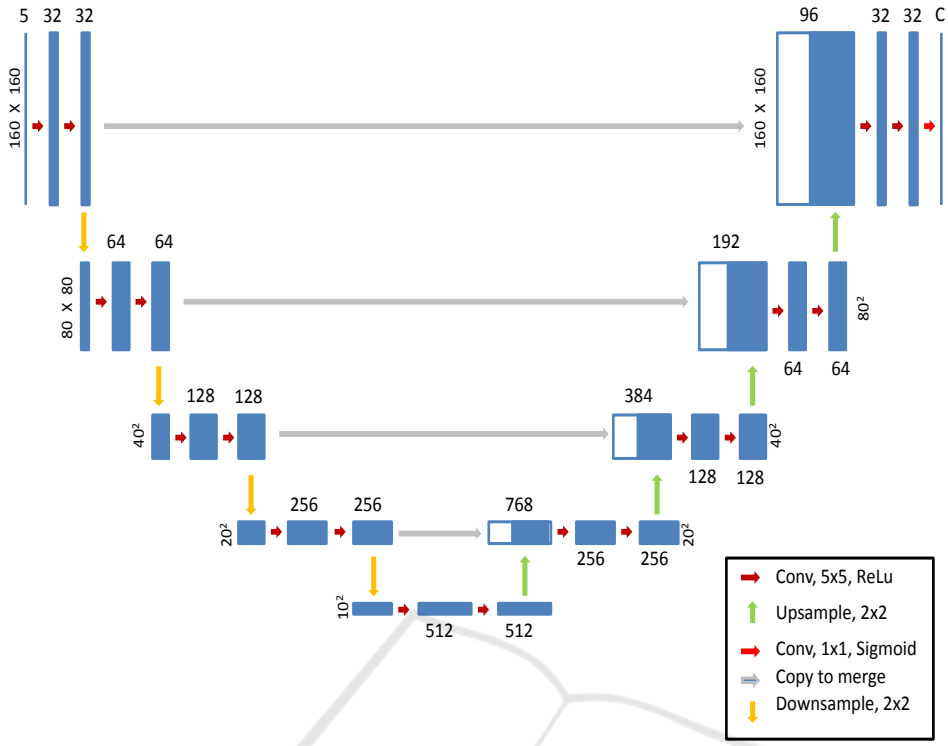


Figure 1: mUnet architecture.

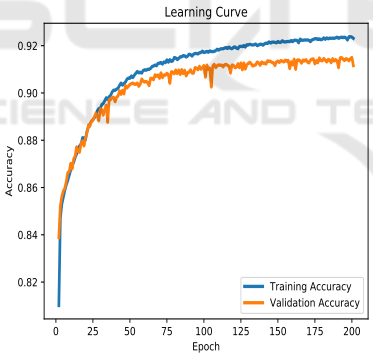


Figure 2: Learning curve.

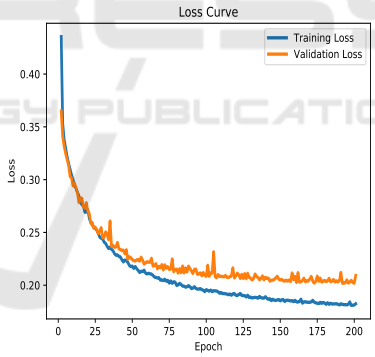


Figure 3: Loss curve.

vectors. The different feature vectors are clustered according to a labeled priority setting (1.Water, 2.Habitation, 3.Vegetation, 4.Open) in order to produce a singular vector containing lowest priority feature at the lowest layer and highest priority feature at the top most layer. The shape regularization is done using ArcGIS (simplify polygon tool) by eliminating undesirable artifacts in polygons geometry. The vector is then converted into thematic raster based on the priority set labels resulting in final LULC map generation. Figure 4 shows the output binary images of the mUnet model for all the four classes

5 RESULTS AND ANALYSIS

The result evaluation is done by calculating kappa coefficient and by calculating the overall classification accuracy. The classification accuracy is assessed by comparing the ground truth thematic map with the classified map generated by the mUnet model. A systematic comparison of our model is done with two deep convolutional networks (Unet, FCN). A FCN architecture consisting of first 5 convolutional blocks similar to VGG-16 was implemented for comparison, however the fully connected layers were re-

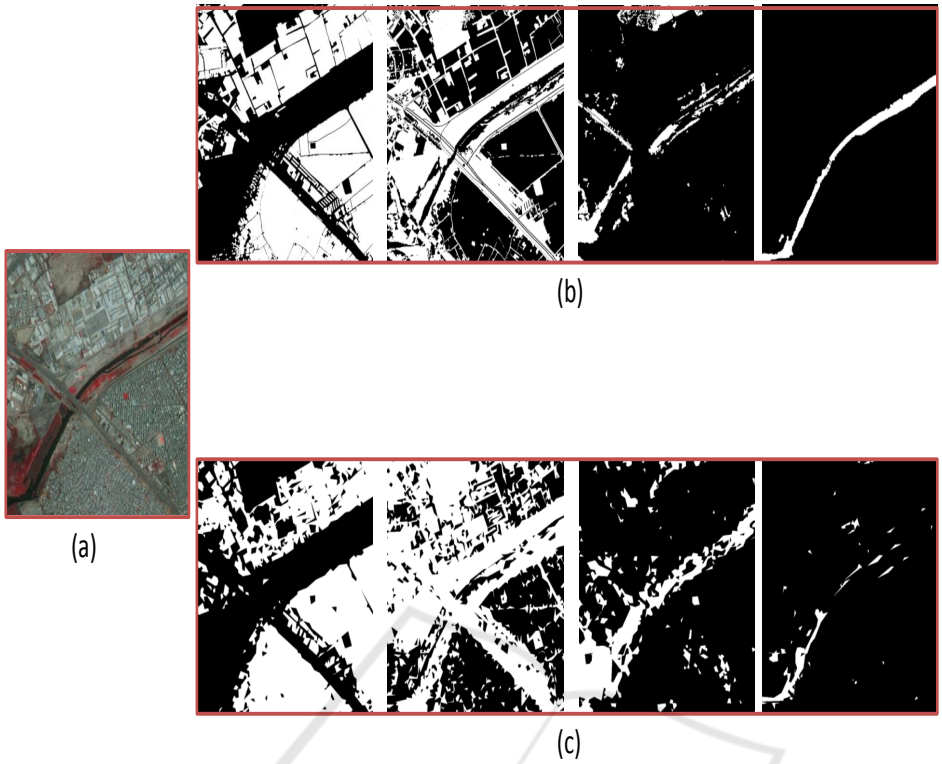


Figure 4: Karachi study area: (a) Satellite image, (b) (left to right) Ground truth for classes habitation, open, vegetation, water, (c) (left to right) predicted map for classes habitation, open, vegetation, water.

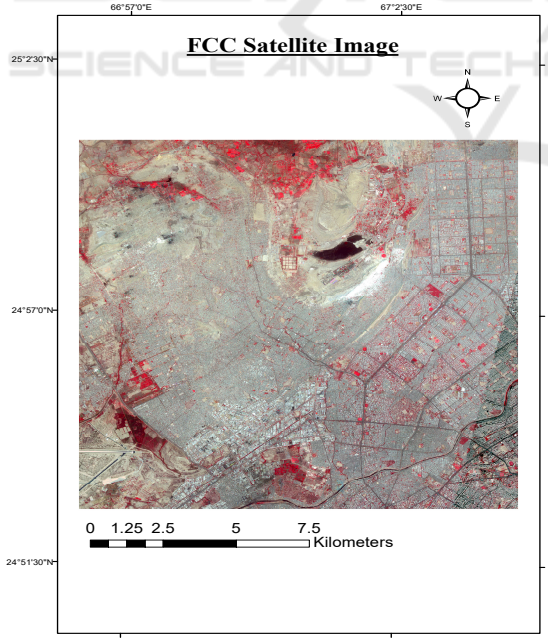


Figure 5: FCC satellite imagery of study area.

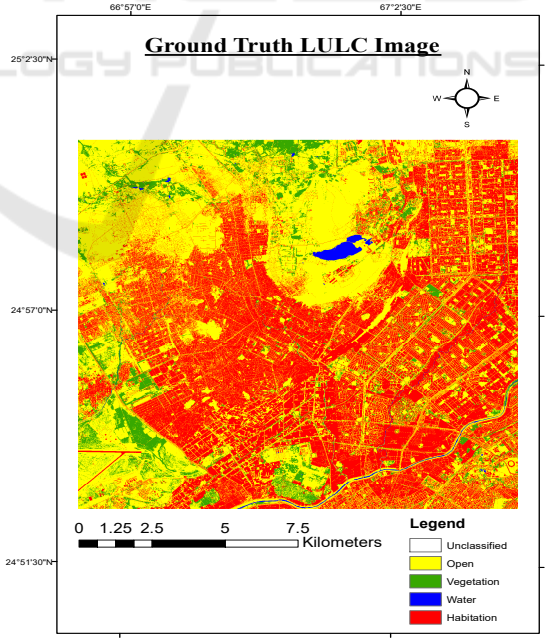


Figure 6: Thematic ground truth map.

placed by convolutional layers. Both the models (Unet, FCN) were implemented using Keras with theano as back end. Table 1 shows the comparison of all three models using three evaluation criterion that are: number of trainable parameters, kappa coefficient, overall accuracy. The model using the least pa-

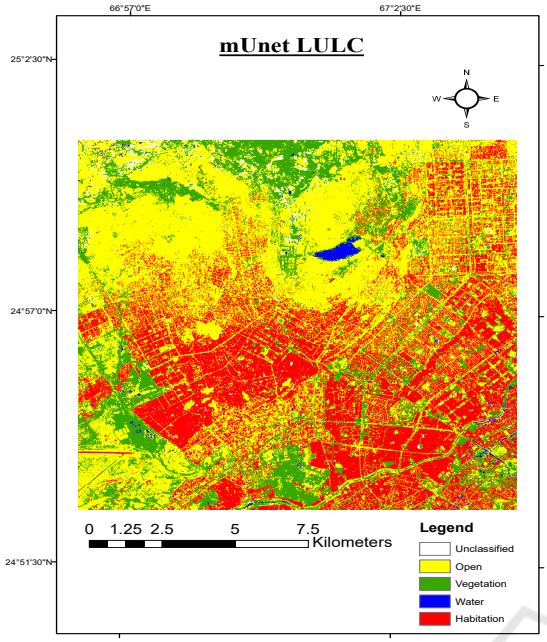


Figure 7: Output of mUnet model.

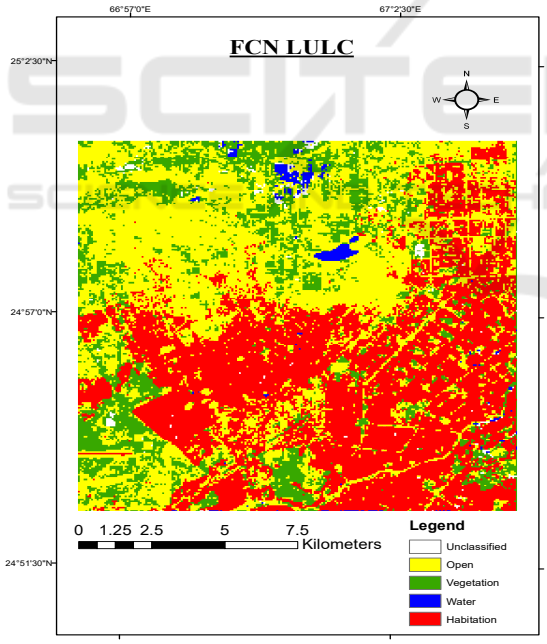


Figure 8: Output of FCN model.

rameters for training and higher value of kappa coefficient and overall accuracy is considered as the best. Figure 2, 3 also showcase the learning curve and the loss curves for our model.

Kappa coefficient measures the agreement between truth values and classification. A kappa value of 1 represents a perfect agreement while a value of 0 represents no agreement. Kappa coefficient is measured as:

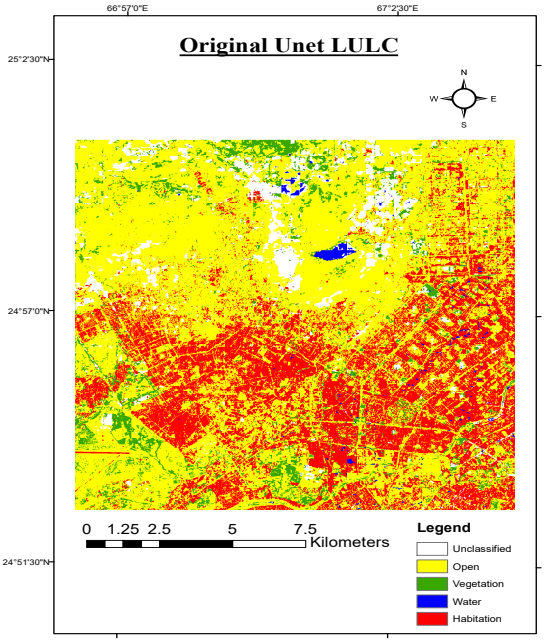


Figure 9: Output of Unet model.

Table 1: Comparison between different segmentation networks.

	$Parameters	*kappa*	*OverallAccuracy*
FCN	134,281,029	.3908	62.446%
Unet	31,379,205	.3558	62.1077%
mUnet	21,791,109	.5501	70.2857%

$$k = \frac{N\sum_{i=1}^{C} m_{ii} - \sum_{i=1}^{C} G_i D_i}{N^2 - \sum_{i=1}^{N} G_i D_i} \quad (2)$$

where N is the total number of pixels and C is the total number of classes. m_{ii} represents the number of pixels of class i classified as class i. G_i are the total number of pixels in ground truth belonging to class i. D_i is the total number of pixels predicted in class i. Accuracy for class c is computed as m_{ii}/G_i.

It is observed that despite having less number of parameters mUnet outperforms other architectures in terms of kappa coefficient and overall accuracy.

Figure 8, 9, 7 show the spatial comparison of LULC classified map generated using all three models. As seen in the figure 7, mUnet performs better than the other two models with providing more accurate results when compared with the ground truth image (figure 6). Original Unet has misclassified open area with habitation and vegetation while some of the areas remain unclassified as seen in figure 6, 9. On

the other hand classification done by the FCN model as seen in figure 6, 8 has reduced the unclassified area but has misclassified pixels belonging to open, vegetation and water classes. When comparing mUnet with other two models the classification results are more accurate with respect to the ground truth (figure 6). Hence we can conclude that our proposed model has performed better for LULC classification than the previously defined models.

6 CONCLUSIONS

This paper has showcased the use of convolutional neural networks for predicting different land use classes from satellite imagery. We study the topographical land use distribution of Karachi using a new CNN architecture, **mUnet**. We benchmark mUnet with other state-of-the-art models such as FCN, Unet. The experimental results demonstrate that our novel approach outperforms FCN and Unet. It is observed that mUnet achieves a higher overall accuracy and kappa coefficient than the other two models. Additionally, mUnet has the advantage that it uses less number of trainable parameters. It can further be taken into consideration that while evaluating our model we utilized a dataset pre-trained to a developing country counterpart to the other studies conducted in the area which have only utilized developed country datasets. As future directions of this work we plan to address the problem of LULC by utilizing more classes in our model, improve and further curate the Karachi dataset and extend this type of analysis to more developing cities across other geographical locations.

REFERENCES

Albert, A., Kaur, J., and Gonzalez, M. C. (2017). Using convolutional networks and satellite imagery to identify patterns in urban environments at a large scale. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 1357–1366.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Iandola, F. N., Moskewicz, M. W., Ashraf, K., Han, S., Dally, W. J., and Keutzer, K. (2016). Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5MB model size. *CoRR*, abs/1602.07360.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105.

Kussul, N., Lavreniuk, M., Skakun, S., and Shelestov, A. (2017). Deep learning classification of land cover and crop types using remote sensing data. *IEEE Geoscience and Remote Sensing Letters*, 14:778–782.

Mnih, V. (2013). Machine learning for aerial image labeling.

Papadomanolaki, M., Vakalopoulou, M., Zagoruyko, S., and Karantzalos, K. (2016). Benchmarking deep learning frameworks for the classification of very high resolution satellite multispectral data. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, III-7:83–88.

Ronneberger, O., P.Fischer, and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *LNCS*, pages 234–241.

Tong, X.-Y., Lu, Q., Xia, G.-S., and Zhang, L. (2018). Large-scale land cover classification in gaofen-2 satellite imagery. *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 3599–3602.

Torres, M. T., Perrat, B., Alfano, V., Goulding, J., and Valstar, M. (2017). Automatic pixel-level land-use prediction using deep convolutional neural networks.

SCIENCE AND TECHNOLOGY PUBLICATIONS

arXiv:1912.10752v1 [cs.LG] 23 Dec 2019

Learn-able parameter guided Activation Functions

S.Balaji[1][1], T.Kavya[1][2], and Natasha Sebastian[2][2]

[1] HCL Technologies
[2] Delhi Technological University

Abstract. In this paper, we explore the concept of adding learn-able slope and mean shift parameters to an activation function to improve the total response region. The characteristics of an activation function depend highly on the value of parameters. Making the parameters learn-able, makes the activation function more dynamic and capable to adapt as per the requirements of it's neighboring layers. The introduced slope parameter is independent of other parameters in the activation function. The concept was applied to ReLU to develop Dual Line and Dual Parametric ReLU activation function. Evaluation on MNIST and CIFAR10 show that the proposed activation function Dual Line achieves top-5 position for mean accuracy among 43 activation functions tested with LENET4, LENET5 and WideResNet architectures. This is the first time more than 40 activation functions were analyzed on MNIST and CIFAR10 dataset at the same time. The study on the distribution of positive slope parameter β indicates that the activation function adapts as per the requirements of the neighboring layers. The study shows that model performance increases with the proposed activation functions.

Keywords: Activation function· Dual Line · DP ReLU

1 Introduction

The activation functions used across the layers of deep neural networks play a significant role in the ability of the whole network to achieve good performance. Though each layer of a deep neural network can have different requirements, the convention is to use the same activation function at each output of a layer. Therefore using a good activation function suitable for every output node in the layer is essential.

Rectified Linear Units (ReLU) [14] and its variants such as Leaky ReLU [11], Parametric ReLU (PReLU)[6] are the most commonly used activation functions due to their simplicity and computational efficiency. The introduction of learn-able parameter in the negative axis for PReLU increased the overall response region. Though extra parameters are introduced in PReLU activation function, the number of parameters added due to PReLU is negligible compared to the total number of parameters in the network. Another activation function that makes use of learn-able parameters is Parametric ELU(PELU) [23]. Learn-able

parameters in PELU activation function adopted characteristics as per the requirements of the training stage. In the case of PELU, the positive axis parameter is dependent as it is defined as the ratio of parameters used to alter exponential decay and saturation point. In this paper, we introduce two activation functions, with a positive slope parameter which is independent of other parameters, allowing it to dynamically update.
In Flexible ReLU [18] and General ReLU, mean shift parameters were introduced to shift the mean activation close to zero. The Dual Line activation function can be viewed as a combination of DP ReLU with learn-able mean shift parameter. The better performance of Dual Line compared to DP ReLU clearly indicates the impact of the mean shift parameter.

The rest of the paper is organized as follows. Section 2 describes our proposed activation function and section 3 deals with their properties. Section 4 describes the steps to be carried out to extend the proposed concept to other activation functions. Experimental analysis and performance evaluation are described in section 5 and section 6 respectively. Results and discussion are detailed in section 7 and 8 respectively. The paper is concluded in section 9.

2 Proposed Activation Functions

2.1 Dual Parametric ReLU (DP ReLU)

DP ReLU is a new variant of ReLU with a learn-able slope parameter in both axes. The difference between Parametric ReLU (PReLU) and DP ReLU is the usage of learn-able slope parameter in the positive axis. For slope parameters (α and β), DP ReLU is defined as

$$X = \begin{cases} \alpha * x, & \text{if } x < 0 \\ \beta * x, & \text{if } x > 0 \end{cases} \tag{1}$$

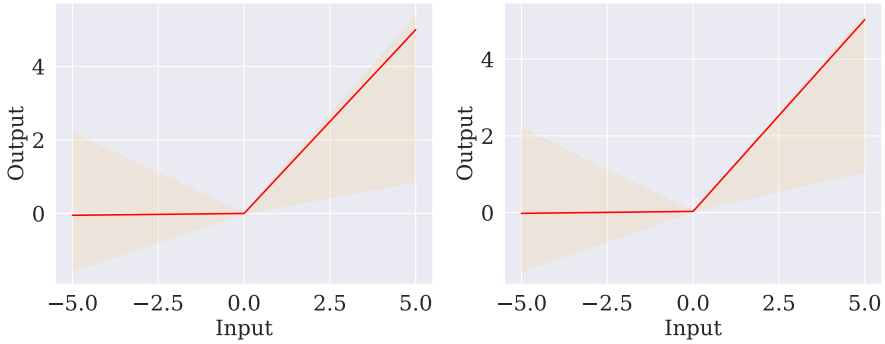
The negative slope parameter α is initialized with a value of 0.01 as in PReLU and Leaky ReLU. The positive slope parameter β is initialized with a value of 1.

2.2 Dual Line

Dual Line is an extension of the DP ReLU activation function. Learn-able slope parameters are multiplied to both axes and the mean shift parameter is added. The resultant activation function resembles the line equation in both axes. For slope parameters (α and β) and mean shift parameter (m), Dual Line is defined as

$$X = \begin{cases} \alpha * x + m, & \text{if } x < 0 \\ \beta * x + m, & \text{if } x > 0 \end{cases} \tag{2}$$

Mean parameter is initialized with a value of -0.22 by adding the mean shift (-0.25) and threshold (0.03) parameters used in TReLU[10].



(a) DP ReLU Activation Function (b) Dual line Activation Function

Fig. 1. Plot of DP ReLU and Dual Line activation function for different values of learn-able parameter. The values were obtained from WideResNet models trained on CIFAR10. Red line indicates the default initialization state. The filled region indicates the overall response region of the activation function, which is obtained by finding the min and max response curves observed across the network for the activation function.

3 Properties of DP ReLU and Dual Line

3.1 Independent

Both the slope parameters are independent of other parameters and act directly on the input without any constraints as shown in Equ. (1) and Equ. (2).

3.2 Large response region

As shown in Fig. 1, the learn-able parameters can take different values, so the proposed activation function has a larger response region compared to the variants without learn-able parameters.

3.3 Slope parameter in positive axis

The value of $\beta >$ x results in boosting the activation and $\beta <$ x results in attenuation of the activation. The final value of β in the model depends on the position of the activation function with respect to other layers.

3.4 Mean shifting due to mean shift parameter

As per Fisher optimal learning criteria, we can reduce the undesired bias shift effect by centering the activation at zero or by using activation with negative values [1]. Unit natural gradient can be achieved by pushing mean activation close to zero. This reduces the risk of over-fitting and allows the network to learn faster [2],[4]. The mean shift parameter in Dual Line activation function aids to push the mean activation towards zero.

3.5 Computation requirements

Using complex mathematical operations in the activation function increases compute time and memory requirement for training and inference. The absence of exponential or division makes ReLU and its variants faster[16]. ISRLU uses inverse square roots instead of exponentials as they exhibit 1.2X better performance in Intel Xeon E5-2699 v3 (Haswell AVX2) [2]. The proposed activation functions does not have complex mathematical operations. The only bottleneck in compute time during training is due to the inclusion of learn-able parameters.

4 Extending the concept to other activation functions

Most activation functions have an unbounded near-linear response in the Ist quadrant. The concept of adding learn-able parameter in the positive axis and mean shift parameter can be extended to other activation functions.

An existing activation function (G) can be modified by treating it as a piecewise function and replacing the characteristics for x $>$ 0. For value of x $>$ 0, the function can be defined as input multiplied by learn-able slope parameter and it remains the same elsewhere.

For an activation function G defined as follows,

$$X = G(x) \tag{3}$$

The proposed concept can be applied as follows

$$X = \begin{cases} G(x) + m, & \text{if } x < 0 \\ \beta * x + m, & \text{if } x > 0 \end{cases} \tag{4}$$

5 Experimental analysis

5.1 Data Analysis

MNIST [13], Fashion MNIST [5], CIFAR10 [3], CIFAR100 [3], ImageNet 2012 [8], Tiny ImageNet [22], LFW [9], SVHN [21], and NDSB [15] are the computer vision datasets used for analyzing activation functions. We are carrying out experiments with MNIST and CIFAR10 in this paper, as they are the most frequently used.

5.2 Experimental setup

PyTorch deep-learning library was used for the experiments [17]. Adam optimizer and Flattened cross-entropy loss are used. Learning rate was estimated using learning rate finder [19]. The max and min values of learning rate across multiple runs are presented, which can be an indicator for the range of values the model prefers.

With hyper-parameters and all other layers kept constant, we modified the activation function to analyze whether the activation function aids the network

to learn faster and achieve better accuracy in minimal epochs. For each activation function, we ran five iterations on each of the datasets. Computational speedup required for analyzing 43 activation functions was achieved using mixed-precision training [12].

5.3 MNIST - LENET5 and LENET4

LENET5 comprises of 2 convolutional layers followed by 3 linear layers. LENET4 comprises of 2 convolutional layers followed by 2 linear layers. Both LENET networks do not have batch normalization layers.

5.4 CIFAR10 - WideResNet

The hyper-parameters were based on fast.ai submission for the DAWNBench challenge [20]. The normalized data were flipped, and random padding was carried out. The training was carried out with 512 as batch size for 24 epochs and learning rate estimated as per the learning rate estimator. Mixup data augmentation was carried out on the data [24].

6 Performance evaluation

Metrics such as accuracy, top-5 accuracy, validation loss, training loss and time are estimated for the 3 networks. WideResNet contains batch norm layers and the LENET network does not. The impact of batch normalization layer would be one of the factors to consider between these architectures. The main analysis parameter is mean accuracy across 5 runs.

7 Results

The following sections discuss the results and analysis of training LENET5 and LENET4 on the MNIST dataset and WideResNet on the CIFAR10 dataset.

7.1 MNIST LENET5

Dual Line achieves the 2nd best accuracy and best mean accuracy. DP ReLU achieves 18th and 19th in accuracy and mean accuracy respectively. Top accuracy is observed in GELU [7].
Dual Line achieves 2nd and 3rd rank w.r.t mean train and validation loss. DP ReLU achieves 18th and 20th in mean train and validation loss respectively.

NAME	Loss		Mean Loss		Acc	Mean Acc	Learning Rate		Time (s)
	Train	Valid	Train	Valid			Min	Max	
ARiA2	0.5207	0.0591	0.5249	0.0652	0.9876	0.9872	0.0025	0.0275	3.83
Atan	0.5335	0.0739	0.5383	0.0788	0.9859	0.9852	0.0063	0.0076	3.43
BentIdentity	0.5368	0.0702	0.5456	0.0756	0.9880	0.9872	0.0012	0.0021	3.63
BReLU	0.5365	0.0721	0.5450	0.0756	0.9857	0.9845	0.0036	0.0275	3.97
dSiLU	0.5799	0.1001	0.5848	0.1049	0.9837	0.9824	0.0331	0.1738	3.77
ELiSH	0.5118	0.0567	0.5166	0.0617	0.9897	0.9880	0.0025	0.0275	3.73
CELU	0.5019	0.0575	0.5049	0.0597	0.9896	0.9890	0.0025	0.0191	3.60
FTSwishPlus	0.5058	0.0609	0.5172	0.0692	0.9878	0.9865	0.0030	0.0275	3.77
GELU	0.5127	0.0541	0.5142	0.0612	0.9905	0.9878	0.0030	0.0229	3.87
GReLU	0.5293	0.0707	0.5357	0.0731	0.9867	0.9855	0.0030	0.0229	3.67
HardSigmoid	0.5646	0.0944	0.5823	0.1001	0.9832	0.9818	0.0275	0.0479	3.70
ISRLU	0.5005	0.0584	0.5052	0.0595	0.9898	0.9888	0.0025	0.0229	3.67
ISRU	0.5249	0.0736	0.5295	0.0795	0.9865	0.9851	0.0110	0.0158	3.70
LiSHT	0.4930	0.0499	0.5116	0.0565	0.9888	0.9881	0.0025	0.0076	3.67
DP ReLU	0.5102	0.0574	0.5183	0.0663	0.9886	0.9875	0.0025	0.0030	4.00
PELU	0.4931	0.0513	0.4983	0.0567	0.9897	0.9885	0.0025	0.0025	4.00
PoLU	0.5047	0.0600	0.5085	0.0641	0.9897	0.9889	0.0030	0.0036	3.87
RationalTanh	0.5243	0.0724	0.5307	0.0815	0.9874	0.9859	0.0036	0.0052	3.70
RectifiedTanh	0.5439	0.0772	0.5492	0.0855	0.9840	0.9827	0.0132	0.0275	3.63
SiLU	0.5010	0.0538	0.5061	0.0593	0.9894	0.9885	0.0025	0.0191	3.30
SQNL	0.5240	0.0676	0.5262	0.0735	0.9886	0.9866	0.0052	0.0076	3.87
Swish	0.5038	0.0561	0.5113	0.0621	0.9891	0.9883	0.0025	0.0229	3.60
ThresholdedReLU	2.3010	2.3010	2.3011	2.3010	0.1135	0.1135	0.0000	0.0000	3.67
TReLU	0.5155	0.0567	0.5207	0.0642	0.9888	0.9877	0.0025	0.0229	3.97
ELU	0.5014	0.0600	0.5079	0.0617	0.9892	0.9884	0.0025	0.0030	3.33
HardShrink	0.6253	0.1031	0.6325	0.1073	0.9769	0.9765	0.0025	0.0191	3.37
HardTanh	0.5296	0.0695	0.5382	0.0772	0.9872	0.9852	0.0030	0.0044	3.47
LeakyReLU	0.5147	0.0640	0.5206	0.0679	0.9886	0.9879	0.0025	0.0030	3.30
LogSigmoid	0.5923	0.0934	0.5964	0.1004	0.9794	0.9787	0.0132	0.1738	3.47
Dual Line	0.4943	0.0542	0.5002	0.0580	0.9901	0.9895	0.0025	0.0275	4.03
Mish	0.5043	0.0502	0.5117	0.0585	0.9901	0.9884	0.0025	0.0229	3.93
LeakyReLU(0.01)	0.5221	0.0621	0.5302	0.0705	0.9874	0.9865	0.0030	0.0191	3.77
PReLU(0.01)	0.5101	0.0532	0.5156	0.0641	0.9897	0.9881	0.0030	0.0191	3.90
PReLU	0.5107	0.0631	0.5169	0.0657	0.9889	0.9879	0.0025	0.0191	3.47
PReLUC	0.5136	0.0619	0.5156	0.0653	0.9896	0.9883	0.0030	0.0191	3.67
ReLU	0.5244	0.0671	0.5335	0.0737	0.9870	0.9859	0.0030	0.0191	3.37
ReLU6	0.5243	0.0657	0.5295	0.0721	0.9872	0.9868	0.0030	0.0275	3.60
RReLU	0.5263	0.0590	0.5295	0.0643	0.9886	0.9875	0.0030	0.0229	3.30
SELU	0.5079	0.0609	0.5125	0.0642	0.9900	0.9894	0.0025	0.0030	3.37
Softshrink	0.6105	0.0923	1.9630	1.8593	0.9804	0.2869	0.0000	0.1445	3.60
Softsign	0.5406	0.0803	0.5456	0.0872	0.9852	0.9848	0.0110	0.0191	3.73
Tanh	0.5201	0.0695	0.5298	0.0762	0.9862	0.9853	0.0052	0.0132	3.70
Tanhshrink	0.5799	0.0746	0.5905	0.0840	0.9826	0.9813	0.0063	0.0191	3.40
	Low								High

Fig. 2. Results for LENET5 network with different activation functions trained on the MNIST dataset. The lite to dark transition corresponds to low to high values. For loss and time, low values are preferred. For accuracy, high values are preferred. Time refers to average training time per epoch in seconds.

Name	Loss		Mean Loss		Acc	Mean Acc	Learning Rate		Time (s)
	Train	Valid	Train	Valid			Min	Max	
ARiA2	0.9027	0.3845	0.9154	0.4000	0.9105	0.9072	0.0692	0.3631	4.00
Atan	1.3547	0.9845	1.3723	1.0030	0.8091	0.8030	0.3020	0.3631	3.75
BentIdentity	0.7807	0.2483	0.7855	0.2545	0.9449	0.9417	0.0479	0.0832	3.90
BReLU	0.9865	0.4770	1.0585	0.5346	0.8844	0.8706	0.1202	0.5248	3.90
dSiLU	2.2823	2.2742	2.2856	2.2783	0.2042	0.1522	0.7586	1.5849	3.90
ELiSH	0.9135	0.3979	0.9309	0.4082	0.9084	0.9035	0.1000	0.3631	3.85
CELU	0.8782	0.3683	0.8857	0.3769	0.9114	0.9104	0.1202	0.1738	3.55
FTSwishPlus	0.9160	0.3965	0.9368	0.4165	0.9051	0.9009	0.1000	0.2089	4.00
GELU	0.8942	0.3810	0.9027	0.3892	0.9108	0.9091	0.1738	0.3631	4.00
GReLU	0.8977	0.3861	0.9195	0.4070	0.9086	0.9015	0.1202	0.2512	4.00
HardSigmoid	2.2794	2.2697	2.2839	2.2755	0.2098	0.1737	0.0000	0.7586	3.60
ISRLU	0.8855	0.3710	0.8925	0.3788	0.9118	0.9106	0.1202	0.1738	4.00
ISRU	1.9906	1.8473	2.0332	1.9064	0.7035	0.6752	0.7586	0.7586	3.95
LiSHT	0.8144	0.2871	0.8292	0.3051	0.9477	0.9436	0.0331	0.0832	3.95
DP ReLU	0.8384	0.3127	0.8651	0.3379	0.9274	0.9191	0.0575	0.2512	4.00
PELU	0.7169	0.1791	0.7241	0.1848	0.9627	0.9618	0.0191	0.0331	4.00
PoLU	0.8747	0.3665	0.8817	0.3744	0.9158	0.9128	0.1445	0.1738	4.00
RationalTanh	1.0322	0.5980	1.0398	0.6059	0.8703	0.8670	0.1000	0.3020	3.70
RectifiedTanh	2.1375	2.0648	2.1748	2.1170	0.7155	0.6346	0.4365	0.9120	3.75
SiLU	0.9037	0.3927	0.9362	0.4144	0.9084	0.9006	0.1000	0.2512	3.60
SQNL	1.5591	1.2356	1.5823	1.2639	0.7738	0.7598	0.3631	0.5248	4.00
Swish	0.9155	0.3943	0.9262	0.4091	0.9059	0.9012	0.0832	0.3631	3.85
ThresholdedReLU	1.5965	1.0087	1.8945	1.5258	0.7938	0.5504	0.2089	0.3631	3.70
TReLU	0.9145	0.3997	0.9285	0.4105	0.9048	0.9027	0.1000	0.3631	4.00
ELU	0.8742	0.3648	0.8879	0.3783	0.9163	0.9093	0.1202	0.2512	3.55
HardShrink	0.8914	0.3853	0.9013	0.3943	0.9036	0.9013	0.0692	0.2089	3.30
HardTanh	0.9844	0.5317	0.9935	0.5401	0.8801	0.8776	0.0832	0.1738	3.40
LeakyReLU	0.9104	0.3907	0.9232	0.4113	0.9108	0.9020	0.1445	0.2512	3.45
LogSigmoid	2.1802	2.0915	2.2231	2.1712	0.5505	0.4447	0.7586	1.0965	3.30
Dual Line	0.8415	0.3101	0.8517	0.3231	0.929	0.9242	0.0832	0.3631	4.00
Mish	0.8942	0.3847	0.9056	0.3923	0.9099	0.9074	0.1202	0.6310	4.00
LeakyReLU(0.01)	0.9103	0.3935	0.9151	0.4011	0.9073	0.9038	0.0832	0.2512	4.00
PReLU(0.01)	0.9075	0.3910	0.9286	0.4133	0.9088	0.9007	0.0832	0.2512	4.00
PReLU	0.9035	0.3871	0.9113	0.3935	0.9079	0.9054	0.0692	0.3631	3.40
PReLUC	0.8965	0.3798	0.9071	0.3932	0.9094	0.9049	0.1202	0.2089	3.45
ReLU	0.8943	0.3775	0.9306	0.4143	0.9124	0.9013	0.0021	1.9055	3.65
ReLU6	0.9293	0.4298	0.9660	0.4796	0.9096	0.8962	0.0692	1.9055	3.45
RReLU	0.8940	0.3757	0.9203	0.4004	0.9098	0.9028	0.1445	0.2089	3.45
SELU	0.8395	0.3402	0.8454	0.3457	0.9247	0.9212	0.0832	0.1738	3.60
Softshrink	0.9069	0.3807	1.2206	0.7472	0.9069	0.8124	0.1445	0.3631	3.50
Softsign	2.0647	1.9518	2.0995	2.0019	0.6998	0.6674	0.7586	1.0965	3.75
Tanh	1.4497	1.1027	1.5161	1.1816	0.7933	0.7725	0.3631	0.4365	3.60
Tanhshrink	2.3021	2.3017	2.3024	2.3022	0.1203	0.1061	0.0000	3.3113	3.45
	Low								High

Fig. 3. Results for LENET4 network with different activation functions trained on the MNIST dataset. The lite to dark transition corresponds to low to high values. For loss and time, low values are preferred. For accuracy, high values are preferred. Time refers to average training time per epoch in seconds.

Name	Loss		MeanLoss		Acc	Mean Acc	LearningRate		Time (s)
	Train	Valid	Train	Valid			Min	Max	
ARiA2	0.6730	0.2046	0.6772	0.2103	0.9459	0.9435	0.0021	0.0044	31.44
Atan	1.3802	1.0106	1.4689	1.1309	0.6565	0.6126	0.0001	0.0002	25.03
BentIdentity	0.9018	0.3904	0.9952	0.4948	0.8847	0.8475	0.0001	0.0010	30.99
BReLU	0.6792	0.2026	0.6872	0.2153	0.9461	0.9417	0.0017	0.0052	26.52
dSiLU	1.2013	0.7770	1.9105	1.7212	0.7456	0.3508	0.0002	3.3113	31.75
ELiSH	0.6933	0.2137	0.6990	0.2210	0.9433	0.9403	0.0017	0.0030	38.08
CELU	0.8654	0.3565	0.9107	0.4002	0.8959	0.8800	0.0004	0.0010	24.31
FTSwishPlus	0.6785	0.2038	0.6825	0.2111	0.9451	0.9418	0.0025	0.0091	29.52
GELU	0.6815	0.2053	0.6988	0.2249	0.9462	0.9384	0.0010	0.0036	36.73
GReLU	0.6801	0.2068	0.6987	0.2265	0.9443	0.9378	0.0014	0.0044	24.04
HardSigmoid	1.1086	0.6637	1.8648	1.9718	0.7913	0.2618	0.0005	2.2909	27.52
ISRLU	0.7812	0.2814	0.7921	0.2902	0.9213	0.9178	0.0010	0.0014	40.20
ISRU	1.1802	0.7664	1.2999	0.9147	0.7487	0.6944	0.0002	0.0006	35.00
LiSHT	0.7050	0.2584	0.9193	0.4894	0.9281	0.8444	0.0004	0.0036	30.50
DPReLU	0.6539	0.2024	0.6767	0.2242	0.945	0.9382	0.0010	0.0036	34.17
PELU	0.6723	0.2117	0.7233	0.2488	0.942	0.9301	0.0003	0.0036	42.18
PoLU	0.8439	0.3334	0.9512	0.4611	0.9023	0.8580	0.0001	0.0014	33.72
RationalTanh	1.4147	1.0576	1.5336	1.2174	0.6389	0.5813	0.0001	0.0002	26.00
RectifiedTanh	0.7340	0.2530	0.7886	0.3046	0.9289	0.9096	0.0012	0.0036	26.00
SiLU	0.7196	0.2224	0.7263	0.2346	0.9416	0.9362	0.0014	0.0044	27.01
SQNL	1.1117	0.6634	1.3628	0.9915	0.7881	0.6636	0.0001	0.0007	39.82
Swish	0.7203	0.2288	0.7250	0.2348	0.9375	0.9354	0.0014	0.0063	28.00
ThresholdedReLU	1.2577	0.8325	1.3098	0.9247	0.7282	0.6980	0.0036	0.0110	24.00
TReLU	0.6869	0.2147	0.7006	0.2273	0.9411	0.9369	0.0014	0.0030	25.02
ELU	0.8945	0.3837	0.9479	0.4448	0.887	0.8651	0.0002	0.0007	24.08
HardShrink	1.4588	1.0938	1.5582	1.2312	0.6223	0.5747	0.0001	0.0002	24.03
HardTanh	1.6641	1.3785	1.6785	1.3980	0.517	0.5116	0.0000	0.0001	24.08
LeakyReLU	0.7039	0.2294	0.7210	0.2413	0.9362	0.9324	0.0010	0.0030	24.01
LogSigmoid	0.8159	0.3091	0.8435	0.3379	0.9129	0.9027	0.0005	0.0025	25.00
DualLine	0.6549	0.2025	0.6599	0.2139	0.9451	0.9422	0.0014	0.0036	35.03
Mish	0.7128	0.2252	0.7337	0.2469	0.9377	0.9308	0.0010	0.0036	29.00
LeakyReLU(0.01)	0.6795	0.2088	0.7042	0.2309	0.9434	0.9366	0.0010	0.0044	24.65
PReLU(0.01)	0.6634	0.2095	0.6699	0.2183	0.9429	0.9405	0.0017	0.0025	25.80
PReLU	0.6578	0.2065	0.6869	0.2280	0.9444	0.9369	0.0008	0.0030	25.51
PReLUC	0.7559	0.2662	0.7941	0.3008	0.9229	0.9126	0.0006	0.0010	25.00
ReLU	0.6900	0.2246	0.7021	0.2334	0.9389	0.9347	0.0014	0.0025	24.08
ReLU6	0.6718	0.2105	0.6923	0.2235	0.9419	0.9384	0.0017	0.0052	24.09
RReLU	0.7469	0.2529	0.7806	0.2817	0.9298	0.9192	0.0007	0.0025	26.00
SELU	0.9964	0.5017	1.1347	0.6773	0.8449	0.7820	0.0001	0.0008	24.18
Softshrink	1.0440	0.5593	1.2974	0.8722	0.8261	0.7061	0.0001	0.0010	24.11
Softsign	0.9965	0.5185	1.2123	0.7969	0.8393	0.7377	0.0001	0.0017	29.20
Tanh	1.3810	1.0148	1.4242	1.0742	0.6558	0.6333	0.0001	0.0002	24.37
Tanhshrink	1.4517	1.0636	1.7615	1.7069	0.638	0.3812	0.0000	0.0479	26.14
	Low								High

Fig. 4. Results for WideResNet with different activation functions trained on the CIFAR10 dataset. The lite to dark transition corresponds to low to high values. For loss and time, low values are preferred. For accuracy, high values are preferred. Time refers to average training time per epoch in seconds.

7.2 MNIST LENET4

Dual Line secure 4th and 5th rank w.r.t accuracy and mean accuracy. DP ReLU secures 5th and 6th position in accuracy and mean accuracy. PELU achieves the best performance in each of the metrics.
Dual Line achieves 5th and 4th in mean train and validation loss. DP ReLU secures 6th and 5th in mean train and mean validation loss.

7.3 CIFAR10 WideResNet

DP ReLU and Dual Line achieve 9th and 2nd best mean accuracy. Aria2 achieves the best mean accuracy of 0.9435. The highest accuracy value was observed with GELU. Dual Line and DP ReLU achieve 4th and 6th best accuracy. Best mean top-5 accuracy is observed in General ReLU. Dual Line and DP ReLU achieve 5th and 15th in mean top-5 accuracy.
DP ReLU and Dual Line achieve the best and 2nd best in Train and Validation loss. Dual Line and DP ReLU secure best and 3rd best mean train loss. 3rd and 8th best in validation loss for Dual Line and DP ReLU respectively.
DP ReLU performance decrease as we start to increase the number of linear layers in the model, which can be seen by comparing its performance in LENET4 and LENET5 models.

8 Discussion

8.1 Learning Rate Analysis

The learning rate estimated varies w.r.t activation used. Learning rates were estimated to indicate the range of values an activation function prefers for a dataset and are shown in the Fig. 2, Fig. 3 and Fig. 4.

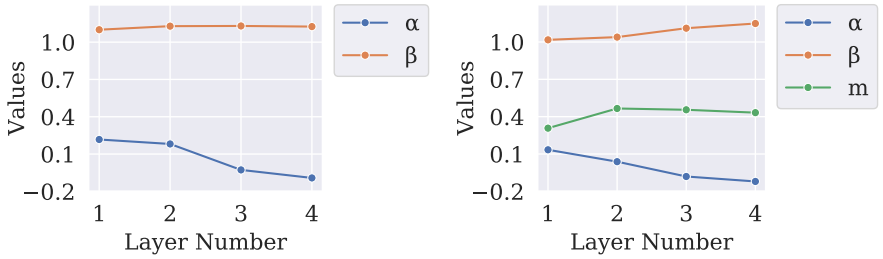
Table 1. Analysis of learning rate values observed for each of the datasets across 5 runs. The 'Overall' row indicates the overall maximum and minimum value observed for a dataset and name of the corresponding activation function.

Activation	LENET5		LENET4		WideResNet	
Function	Min	Max	Min	Max	Min	Max
DP Relu	2.5E-03	3.0E-03	5.7E-02	2.5E-01	1.0E-03	3.6E-03
Dual Line	2.5E-03	2.7E-02	8.3E-02	3.6E-01	1.4E-03	3.6E-03
Overall	1.0E-06	1.7E-01	1.0E-06	3.3E+00	1.0E-06	3.3E+00
	ThresholdedReLU	LogSigmoid	Tanshrink	Tanshrink	Tanshrink	dSiLU

Table 1 shows the maximum and minimum values observed for each of the datasets. Overall, Dual Line prefers the range as 1.4 E-03 to 0.363 and DP ReLU prefers 1E-03 to 0.251. Higher learning rates are observed in LENET4 compared to WideResNet and LENET5.

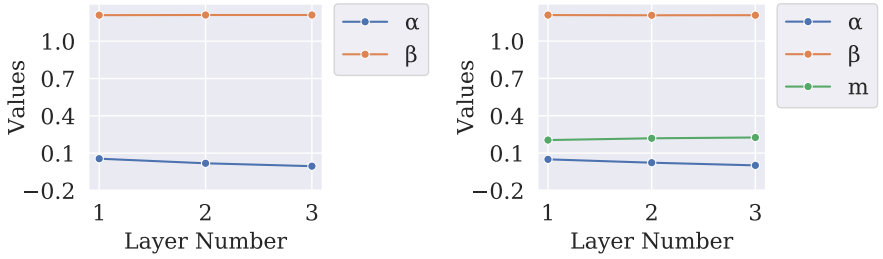
8.2 Parameter Value Analysis

The value of the learn-able parameters used in the activation function decides the response region, thereby the characteristics of the activation function. The neighboring blocks have an impact on the activation function as it receives input from them. The activation function requirement may vary based on the position of activation function within a block of a network, which can be analyzed by checking the parameter distribution of the activation function. As LENET5 and LENET4 models don't have repeating blocks, we are not able to perceive any patterns as shown in Fig. 5 and Fig. 6.



(a) DP ReLU activation function (b) Dual Line activation function

Fig. 5. Value of parameters w.r.t position of the activation function from top to bottom of LENET5 model

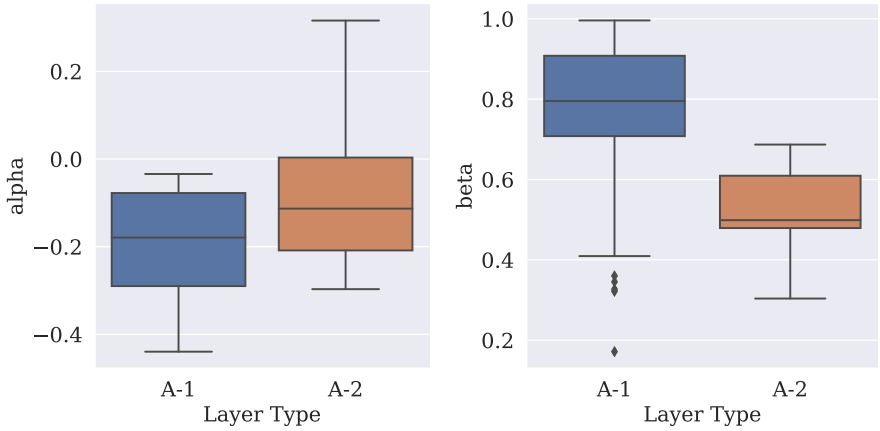


(a) DP ReLU activation function (b) Dual Line activation function

Fig. 6. Value of parameters w.r.t position of the activation function from top to bottom of LENET4 model

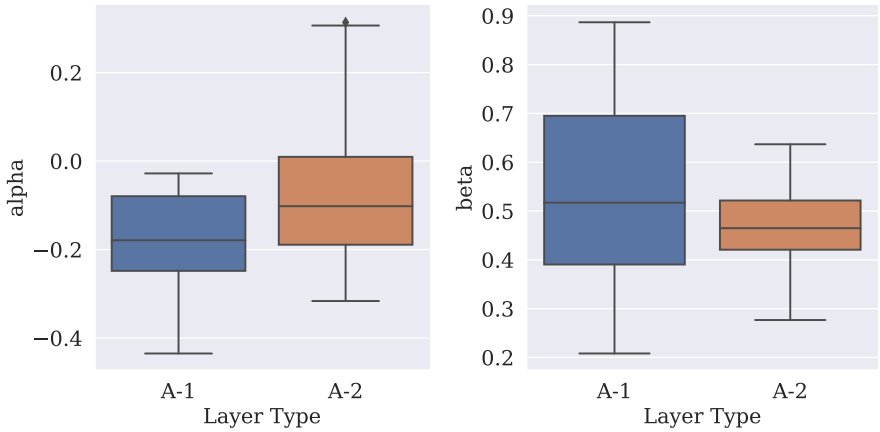
In WideResNet, activation function occurs twice in each of the nine repeating blocks of the network. The distribution of parameter values w.r.t each of these blocks is analyzed to view the relationship existing between activation functions occurring within a same block, as shown in Fig. 9a and Fig. 9b. The value of β of the 1st activation function within a block is larger than the 2nd activation function in the blocks(B1-B8). The first block (B0) which is close to the input, does not exhibit this pattern.

The box plot of α value indicates the marginal difference in distribution within a block as shown in Fig. 7a. The distribution of β value differs a lot within a block as shown in Fig. 8b. The outliers present in the box plot are due to the β values from block (B0). This indicates that the activation function requirements differ within a block of a network.



(a) Distribution of α parameter (b) Distribution of β parameter

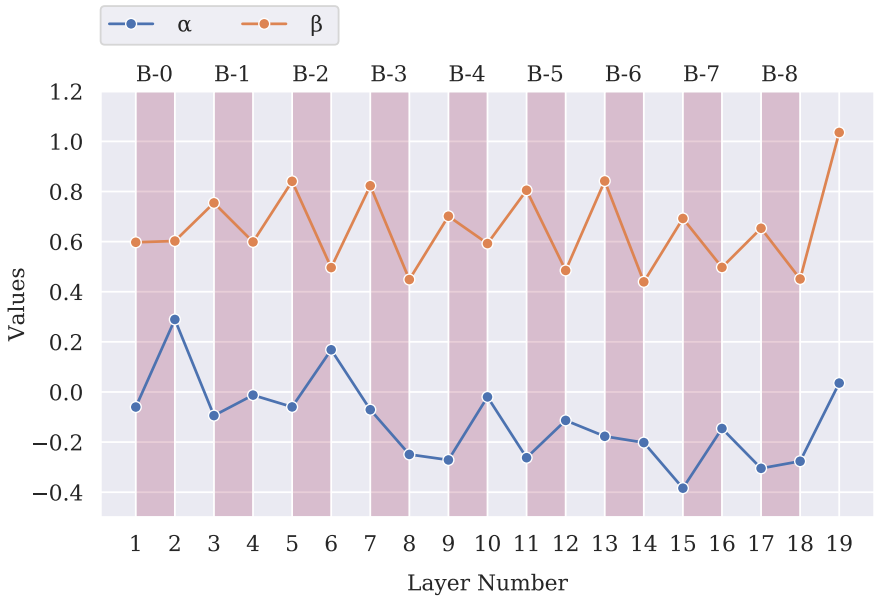
Fig. 7. Box plot of parameters of DP ReLU activation within each block. A-1 and A-2 represent the first and second activation function within each block of WideResNet.



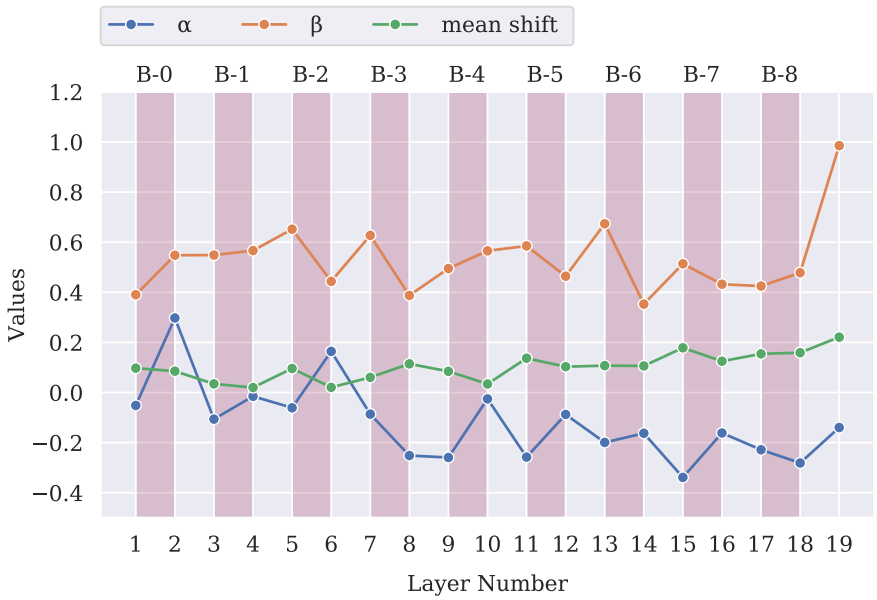
(a) Distribution of α parameter (b) Distribution of β parameter

Fig. 8. Box plot of parameters of Dual Line activation within each block. A-1 and A-2 represent the first and second activation function within blocks of WideResNet.

From Fig. 9b, we can observe that the α parameter does not exhibit any significant pattern, the mean shift parameter is nearly constant and β values differ



(a) Value of α and β parameters of DP ReLU trained on CIFAR10



(b) Value of α, β and mean shift parameters of Dual Line trained on CIFAR10

Fig. 9. Distribution of parameter values for proposed activation functions across blocks of WideResNet. X - axes represent the position of the activation function from top to bottom of the network. Red vertical blocks (B-0 to B-8) represent each of the blocks.

marginally for Dual Line. The marginal difference in β distribution can be seen clearly in Fig. 8b. The block-level parameter distribution pattern indicates the final distribution of the parameters and helps us to understand the requirements of activation function at block level. As we have used learn-able parameters, the proposed activation function can adapt as per the requirement posed by its neighboring layers.

8.3 Performance on fast.ai Leaderboards

The proposed activation function Dual Line broke 3 out of 4 fast.ai leaderboards for the image classification task. It was able to achieve more than 2% percent improvement in accuracy in two leaderboards. Parameter distribution pattern within blocks was observed in XResNet-50 models trained for the above 4 leaderboards.
In our current work, we have analyzed the concept of learn-able slope parameter for the positive axis and mean shift parameter and have shown the performance benefit of the same. Our future work deals with reducing the computation time taken during training.

9 Conclusion

The novel concept of adding a learn-able slope and mean shift parameter is introduced in this paper. Overall, our experiments indicate the performance benefit of the proposed concept. The concept can be added to other activation functions with ease for performance boost. As the paper captures the activation function requirement at the block level, the proposed concept can be used as a supporting guideline for developing new activation functions for computer vision.

References

1. ichi Amari, S.: Natural gradient works efficiently in learning. Neural Computation **10**, 251–276 (1998)
2. Carlile, B., Delamarter, G., Kinney, P., Marti, A., Whitney, B.: Improving Deep Learning by Inverse Square Root Linear Units (ISRLUs). arXiv e-prints arXiv:1710.09967 (Oct 2017)
3. CIFAR-10 and CIFAR-100 datasets, `https://www.cs.toronto.edu/~kriz/cifar.html`, Last accessed 28 Aug 2019
4. Clevert, D.A., Unterthiner, T., Hochreiter, S.: Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). arXiv e-prints arXiv:1511.07289 (Nov 2015)
5. Fashion MNIST, `https://www.kaggle.com/zalando-research/fashionmnist`, Last accessed 28 Aug 2019
6. He, K., Zhang, X., Ren, S., Sun, J.: Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. arXiv e-prints arXiv:1502.01852 (Feb 2015)

7. Hendrycks, D., Gimpel, K.: Gaussian Error Linear Units (GELUs). arXiv e-prints arXiv:1606.08415 (Jun 2016)
8. ILSVRC2012, http://www.image-net.org/challenges/LSVRC/2012/, Last accessed 28 Aug 2019
9. Labeled Faces in the Wild Home, http://vis-www.cs.umass.edu/lfw/, Last accessed 28 Aug 2019
10. lessw2020: Trelu, https://github.com/lessw2020/TRelu, Last accessed 28 Aug 2019
11. Maas, A.L.: Rectifier nonlinearities improve neural network acoustic models. In: JMLR (2013)
12. Mixed Precision Training, https://docs.nvidia.com/deeplearning/sdk/mixed-precision-training/index.html#mptrain, Last accessed 28 Aug 2019
13. THE MNIST DATABASE, http://yann.lecun.com/exdb/mnist/, Last accessed 28 Aug 2019
14. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines, https://www.cs.toronto.edu/~fritz/absps/reluICML.pdf, Last accessed 28 Aug 2019
15. National Data Science Bowl, https://www.kaggle.com/c/datasciencebowl/data, Last accessed 28 Aug 2019
16. Nwankpa, C., Ijomah, W., Gachagan, A., Marshall, S.: Activation Functions: Comparison of trends in Practice and Research for Deep Learning. arXiv e-prints arXiv:1811.03378 (Nov 2018)
17. PyTorch, https://pytorch.org/, Last accessed 28 Aug 2019
18. Qiu, S., Xu, X., Cai, B.: FReLU: Flexible Rectified Linear Units for Improving Convolutional Neural Networks. arXiv e-prints arXiv:1706.08098 (Jun 2017)
19. Smith, L.N.: Cyclical Learning Rates for Training Neural Networks. arXiv e-prints arXiv:1506.01186 (Jun 2015)
20. Dawnbench, https://dawn.cs.stanford.edu/benchmark/#cifar10
21. The Street View House Numbers (SVHN) Dataset, http://ufldl.stanford.edu/housenumbers/, Last accessed 28 Aug 2019
22. Tiny ImageNet Visual Recognition Challenge, https://tiny-imagenet.herokuapp.com/, Last accessed 28 Aug 2019
23. Trottier, L., Giguère, P., Chaib-draa, B.: Parametric Exponential Linear Unit for Deep Convolutional Neural Networks. arXiv e-prints arXiv:1605.09332 (May 2016)
24. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz: mixup: Beyond Empirical Risk Minimization. arXiv e-prints arXiv:1710.09412 (Oct 2017)

Making Smart Homes Smarter: Optimizing Energy Consumption with Human in the Loop

Mudit Verma
Computer Science & Engineering
Delhi Technological University
New Delhi, India
mudit.verma2014@gmail.com

Siddhant Bhambri
Computer Science & Engineering
Delhi Technological University
New Delhi, India
siddhant_bt2k16@dtu.ac.in

Arun Balaji Buduru
Computer Science & Engineering
Indraprastha Institute of Information Technology
New Delhi, India
arunb@iiitd.ac.in

arXiv:1912.03298v1 [cs.AI] 6 Dec 2019

***Abstract*—Rapid advancements in the Internet of Things (IoT) have facilitated more efficient deployment of smart environment solutions for specific user requirement. With the increase in the number of IoT devices, it has become difficult for the user to control or operate every individual smart device into achieving some desired goal like optimized power consumption, scheduled appliance running time, etc. Furthermore, existing solutions to automatically adapt the IoT devices are not capable enough to incorporate the user behaviour. This paper presents a novel approach to accurately configure IoT devices while achieving the twin objectives of energy optimization along with conforming to user preferences. Our work comprises of unsupervised clustering of devices data to find the states of operation for each device, followed by probabilistically analysing user behavior to determine their preferred states. Eventually, we deploy an online reinforcement learning (RL) agent to find the best device settings automatically. Results for three different smart homes data-sets show the effectiveness of our methodology. To the best of our knowledge, this is the first time that a practical approach has been adopted to achieve the above mentioned objectives without any human interaction within the system.**

***Index Terms*—Internet of Things; Unsupervised Clustering; Markov Decision Processes; Reinforcement Learning**

I. Introduction

The increasing involvement of smart devices in our lives has necessitated us to come up with better coordination and management strategies for increased efficiencies and Quality of Service (QoS). This increase in the number of devices has resulted in higher energy consumption by households and has led to the problem of energy optimization. Many organizations are investing in extensive research to come up with strategies to make utilization of energy as efficient as possible. Smart Homes or Smart Devices is one such area that makes use of Information & Communication Technologies for finding solutions to such environmental issues. With the help of embedded intelligence, such devices can target the preferences of a user during everyday life. However, cogency of such results is critical [1], [2]. With the installation of smart meters along with these devices at homes, we can find effective ways to provide users with a visualization of their energy consumption pattern in a way they can comprehend [3]. Furthermore, we can even find ways to make intelligent systems that can recommend users or act on their behalf, based on their past power consumption behavior, to achieve the goal of minimizing electricity wastage. The objective here needs to be to prevent any unnecessary consumption of energy, preserving the comfort and productivity of the user. In smart homes, a vast number of heterogeneous appliances, sensors, and actuators inter-operate and provide context information, which in turn, together with user preferences, are used to effectuate a value-added functionality dynamically. A smart home needs to be able to analyse the actions of its occupant, taking into account the context information, to proactively recognize the occupants activity to conserve energy. Therefore, while creating such systems, one must proceed in a way such that the intelligent planners policies in promoting efficient energy usage in households must take into account the users choices and behavior for utmost user satisfaction [4].

Another study of devices reveals a framework where cluster analysis is done in a three-part architecture, starting with maintaining the sequence of individuals activities during the day, showing how long each activity is carried out and ending with finding the clusters for different activity patterns for the group of users [5]. However, the results from this analysis have not been put to any efficient use for the user. Another intelligent planning approach has three entities: the manufacturers who provide the state information, the SESPs (Smart Environment Service Providers) who manually assess the device states and the users get a recommended set of actions to reduce energy consumption [6]. The two frameworks combined cater to the problem, but a lot of manual effort is involved and therefore, faces a lot of challenges regarding the feasibility. Hence, we aim to overcome these challenges by applying a very simple yet effective approach to automate the entire framework from collecting and analysing user data to taking decisions in accordance with the users preferences without involving any manual effort of any kind, once the framework is deployed.

We first elaborate on these and some other challenges along with solutions focused on only some specific parts of the complete problem at hand in Section 2. In Section 3, we explain our approach to adaptively configure smart devices in a smart home through the effective incorporation of probabilistic user behaviour(s). Then, we present a method to automate the complete process of collecting the state information (i.e., finding clusters for different activity patterns), followed by analysing it to define the quality of states (i.e., assess the de-

vice states keeping in mind the user preferences). Eventually, we formulate a RL technique for recommending actions to the user to achieve the set goal of optimal power consumption in Section 4. We conclude by presenting the results of our approach in Section 5 followed by the conclusion in Section 6.

II. Related Work

The increase in computing power in constrained environments have allowed researchers to experiment and come up with new and better approaches to solve some critical problems using the power of artificial intelligence. Hence, there have been some approaches proposed earlier which either have been mainly theoretical in nature or have focused only on one specific part of this problem statement. We present a brief discussion on these methodologies adopted by researchers in the past.

In works that study the behavioural clustering of devices, Adika et.al. present a method where appliances are clustered together based on their hourly energy consumption data [7]. Then a time of use probability distribution is made, and each cluster is given a schedule, making this a job scheduling problem which is solved using dynamic programming. The challenge here is to predict the correct cost of energy for the next hour. Consequently, the scheduling assumes that the error in predicting the cost is minimum. However, the approach does not focus on any method to suggest the user ways to cut down on excessive power consumption.

In yet another approach, the authors introduce a generic way of creating a Peer to Peer (P2P) overlay network using Hydra middle ware [8], [9]. The authors talk about collecting and displaying energy consumption data using several data visualisation techniques. The approach does not have intelligent monitoring or recommendation for the users, which makes us ask the question of whether the data visualisation effects the users understanding of energy consumption or not [3].

Wei et. al. show another general architecture for designing the energy consumption monitoring and energy-saving management system that are IoT based [10]. According to this work, IoT infrastructure has three levels: the bottom-most level collects data from different sensors, the middle level which is the network layer talks about data transmission, and the topmost level processes the collected and transferred data using cloud computing and fuzzy pattern recognition techniques. A 3-component architecture was also introduced to cater to an intelligent household lighting system for efficient energy consumption, including users context information such that the systems behaviour conforms to user satisfaction [11]. Although this approach talks about an intelligent planner, it is limited to the lighting systems and uses a static algorithm (minimum light intensity control) to control the lights. Hence, the approach is not efficient when it comes to dynamic environments where user behaviour is prone to change or in a real-life setting where a set of diverse electrical equipment or appliances are present.

Yau et. al. base their technique on Markov Decision Process (MDP), assuming there are three entities [6]. Firstly, the manufacturer provides the set of device states, secondly a SESP (Smart Environment Service Provider) who analyses the state information and lastly, the MDP planning algorithm which makes a policy of states and actions along with a user who receives these actions suggested by the MDP and acts accordingly in the given environment. The approach assumes a Central Module which collects all the information and does the heavy lifting, in turn, solving the problem of having a constrained environment. The approach involves manual human intervention, firstly as the manufacturer who supplies information on the devices and then in the role of SESP who performs analysis on the data. Also, The results are shown over simulation, and the approach is never used on a real data-set.

The methodology used by Jahn et. al. and Byun et. al. makes us understand the major approaches for collecting data from appliances and hence, helps us move in the direction of intelligent planning [8], [11]. The architecture proposed by Wei et. al. is a general way to move from the stage of data collection to intelligent planning [10]. There are existing approaches that are perfectly capable of achieving optimal energy consumption, taking the assumption that the user behaviour is not dynamic, and the price prediction has minimum error [7], [11]. The approach used by Yau et. al. uses RL and hence, is capable of dealing with a change in user behaviour [6]. However, the role of entities like manufacturer and SESPs can be automated. In this work, we find the quality of states based on the user behaviour, which keeps on updating itself to improve with time. The work presented in this paper is an advancement of the approach discussed in [6].

III. Proposed Methodology

We use a top-down strategy to explain the complete method. First, we give an overview of the whole system and interactions between logical sub-sections. Then we dissect each part and discuss our approach in depth.

A. Flow of Data

According to Figure 1, the model is divided into three segments.

Segment A: This is the first segment in the entire flow of data inside our model. The data is collected from the user, and his/her behaviour/preferences are recorded, which comprise of an *Action* and an *Actuation* as shown in steps 1.a and 1.b of the figure. We define the actual actions taken by the user to change the device settings as *Actuations* whereas the stochastic actions relevant to the RL agent are termed as *Actions*, and these are received by the *Mapper*.

Segment B: In the second segment, the devices' power consumption values are clustered to determine device modes, ${D_i}^{n_i}$ for i^{th} device, where the device has n_i modes of operation. These device modes are again clustered together to form m *domain states* $DS = \{DS_1, DS_2 ... DS_m\}$. Hence, $DS_i = ({D_1}^{n_1}, {D_2}^{n_2}, ... {D_k}^{n_k})$ for k devices. Therefore, D_i

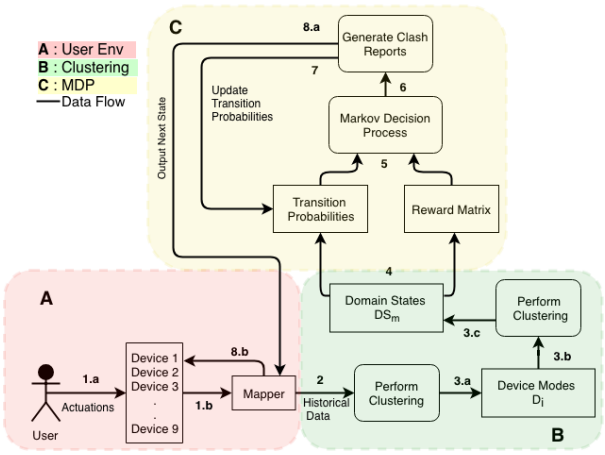


Fig. 1: The flow of data in proposed framework.

signifies device-level representation whereas DS_i is representative of the entire set of devices.

Segment C: The domain states act as an input to the third and last section, i.e. Segment C, where our RL algorithm takes a suitable decision with respect to choosing a domain state optimally. *Transition Probability Matrix* and *Reward Matrix* are two essential components of our RL algorithm which are fed as inputs and these are computed using the Domain States. The RL agent generates an output and compares these with the historical data during the learning phase. It suitably adapts to the users behavior and updates the probabilities of transitioning from one domain state to another.

The **Mapper** entity in segment A of Figure 1 is simulated for the results in our paper as mentioned in sections III-C and V. It is a many-to-many mapping where several actual device settings can correspond to internal device modes and vice versa. This mapping simplifies our approach to search for the appropriate device modes which a user can toggle through a suitable actuation. The live data input source will require recording the actuation signalled by the user and the corresponding device mode action. This mapping may also be clubbed with the clustering algorithms as a feature input so that the clustered nodes also have information about the actuation. Other methods may involve finding the best possible actuation (such as by using costs incurred in shifting device modes) which corresponds to some device action in step 8.b of Figure 1.

B. Two-Step Clustering

As our first step in our proposed approach, we intend to remove the use of SESPs [6] who manually provide device modes & label them as good or bad. This manual labour inhibits the system to achieve complete automation. The two-step clustering refers to using the available raw user data to form device mode clusters followed by using these device modes to form domains state clusters.

We use the data of user actions and perform clustering using time features (hour, month & year) and device power consumption for device modes, where a device is not aware of the existence of another device which ensures independence of device usage, a qualitative feature of appliances/devices in real-life. Hence, device modes are clustered output equivalents of device actuations as discussed in III-A. Domain states, on the other hand, are clustered outputs which take all the devices into consideration, wherein the states are built up by the use of all of the devices modes clustered together.

We use Growing Neural Gas clustering algorithm which yields us with a graph when the input data is projected onto a 2D space. Each vertex of this graph corresponds to a neuron in which input data has been mapped. GNG learns through a combination of updated Kohonen learning approach to adjust the positions of the neurons along with a Competitive Hebbian Learning (CHL) approach for its connections [12]. The number of connected components in this graph represent the total number of clusters.

C. Actions and Actuations

As defined in Segment A of III-A, we further explain these two terms. Initially, we consider a users historical data to supply the actions, i.e., if the current domain state and previous domain state are same, it is taken to be a *STAY* action, otherwise, *MOVE*. For example, let the current domain state for 5 devices be (1,1,0,3,2) and the next domain state for the same 5 devices be either (1,1,0,3,2) or (1,1,0,2,0). In the first scenario, where the device mode values remain the same for all 5 devices, it is said to be a *STAY* action whereas in the second case wherein the device mode values changed for the last two devices, it is said to be a *MOVE* action.

To obtain the actuations, we select 30% of our training data points for which actions differ from the actuations. This implies that the actions and actuations are drawn from a similar distribution. Thus, there are certain data points where a *STAY* action may have resulted in *MOVE* and vice-versa. Hence, these 30% positions signify the underlying stochasticity in the model. We label the time-stamps signifying the actual *MOVE* and *STAY* actions as M and S, and the states representing actuations as m and s.

Now, the model has a probability $P_{dd'}(i \cap j)$ which represents the state transition from domain state d to d when actuation i and action j occurred, where $i \in \{m, s\}$ and $j \in \{M, S\}$. However, these actuations are for training dataset only. For testing data, we find actuations as in section V. Figure 2 shows a sample transition between any two random domain states s_0 and s_1 with the agent currently at s_0. Due to the attached stochasticity, both *MOVE* and *STAY* action can result in a transition from s_0 to s_1 or transition back to the same state s_0. Note that the sum of probabilities for all transitions as *MOVE* or *STAY* is unity, which is implicit.

D. Domains State Types

For understanding the domain states qualitatively, we examined the training data set and divided these domain states into two broad categories- high demand (HD) states and low demand (LD) states. The HD states are the most frequently

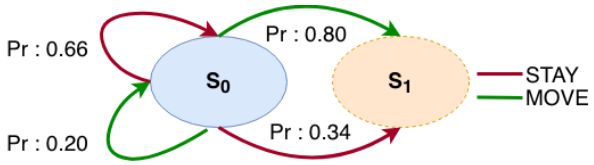


Fig. 2: Possibilities available to a state given the actions: STAY and MOVE

visited states by the user whereas, on the other hand, LD states are the least visited states. Let us say that top percent of the total number of domain states comprise of HD states and the remaining are the LD states.

In a real-life scenario, there may be some domain states which the user may prefer to stick to, irrespective of the output of the RL agent and these domain states could intuitively fall under both the categories. We refer such domain states as strict high demand (SHD) states and say, that the HD states comprise of fix_{HD} percent of SHD states and the remaining HD states are termed as loose high demand (LHD) states for simple reference. Similarly, there may be some infrequently occurring states, i.e. LHD states which the user may prefer to stick to. We refer such domain states as strict low demand (SLD) states and say, that the LD states comprise of fix_{LD} percent of SLD states and the remaining LD states are termed as loose low demand (LLD) states for simple reference.

E. Reinforcement Learning Agent

For our RL agent formulation, we intended to experiment with a simple RL model and hence used the MDP algorithm which provided us satisfactory results. Therefore, this section highlights our MDP formulation for the given problem setting. Henceforth, this provides us with a future opportunity of modelling the current problem statement using more complex RL algorithms and compare the results as shown in the current approach. A Markov Decision Process is a reinterpretation of Markov Chains which includes an agent and a decision-making process [13]. Our MDP is characterised by the following components:

1) State Space: $S = d_0, d_1, ..., d_m$ where d_i represents i^{th} domain state,
2) Initial State: d_0 the first domain state in data-set,
3) Actions: $A = a_0, a_1$ where a_0 is STAY and a_1 is MOVE,
4) Transition Model: $T(s, a, s')$ and
5) Reward Function: $R(s) = -1 * totalPowerConsumptionOfDomainState$.

1) State Space: The clustering of domain states provides an implicit functionality of state space reduction for MDP; therefore, preventing state space explosion. It is possible to perform domain state clustering directly, without the use of first level device mode calculation. However, to enforce the MDP outputs on the system, we must know to which device mode MDP is asking the devices to move, and thus the need for finding device modes.

2) Transition Matrix: The transition matrix returns the probability of reaching the state s if action a is taken in state s. We use the training data set to compute this matrix. These transitions are governed by the reward received by the RL agent on taking a particular action a in state s to reach a new state s. The Reward function $R(s)$ returns a real value every time the agent moves from one state to the other. The agent receives a higher reward (less negative in value) when it reaches one of these desirable states.

3) Updating Transition Matrix: The RL agent gets the current domain state as the input and finds the state the user should be in, which is the output of the MDP function $mdp(s_t)$. While learning on the training data, the MDP agent calculates the number of clashes which take place when the user transitioned to another state while the RL agent recommended a different state. Our MDP algorithm runs online and is, therefore, able to adapt during the testing phase as well. When such a clash occurs, an update of the Transition Matrix involves reducing the probabilities of MDP outputs by a factor $e = 0.1$ and increasing the probabilities of states belonging to the sets SHD and SLD, finally normalizing to maintain the probability sum as unity.

4) MDP Algorithm: The user actions have been modelled via the transition matrix, whereas the need for optimal power consumption is taken care of by the Reward Function. We implement our MDP using policy iteration algorithm where the agent chooses the best state using a policy. Policy iteration is guaranteed to converge to the optimal policy, and it often takes fewer iterations to converge than value iteration. The maximum expected utility principle states that a rational agent should choose its action that maximizes its maximum utility [14]. The utility of a single state is defined as:

$$U(s) = E[\sum_{t=0}^{\infty} \gamma^t R(s_t)] \quad (1)$$

The utility of a state s is correlated with the utility of its neighbor at s' as:

$$U(s) = R(s) + \gamma max \sum_{s'} T(s, a, s')U(s') \quad ; \quad \gamma = 0.9 \quad (2)$$

5) Reward Function: To account for the objective of reducing excessive power consumption, we planned to make use of a simple reward function that caters to our aim of rewarding the RL agent to make a transition to a state where the total power consumption is less than that at the current state. Hence, we want to penalize the domain states with higher power consumption and reward the domain states with lower power consumption. Therefore, we negate the parameter $totalPowerConsumptionOfDomainState$ to imply a more substantial reward for the domain states with lower power consumption.

6) Reward Matrix: The reward matrix is calculated by first finding the domain states, each of which has devices operating in specific modes, thus consuming power. We sum

the consumption values to get a reward value for each domain state. The reward matrix is formed using these calculated reward values for all states. The current reward function penalizes a state for using power. The reward matrix can be created dynamically since with our clustering approach; we get a set of connected components of nodes rather than one single cluster centre. Therefore, a single device mode corresponds to the connected component of the graph. For our results, we use the average power consumption value for all the nodes in a component. However, we may also choose different power consumption values for different nodes to represent the power consumption of the device mode and have a more "intelligen" reward matrix.

F. Method of Validation

Comparing methods for their efficacy in energy optimality is a reasonably easy task by comparing the expected energy consumed. However, this is not so trivial for proving that our approach would work well at preserving human device setting preference. There is no supervised information present as to which states are liked or disliked by the user. Neither can a state lasting over a long time be considered as a desired state nor can an impulsive change from one state to another be labelled as a disliked state. Since application functionality of the devices is not studied; to centrally prevent the need of manual domain experts, we cannot find more likely or less likely preferred states usage-wise solely by the power consumption values. Thus, we require to prove two statements true simultaneously regarding our system to justify the aim of this paper effectively:

1) Reduction in clash rate over time, primarily for HD states.
2) Decrease in overall power consumption.

We define Clash Rate as the number of incorrect next state predictions made by the MDP for the actual future (next time step from the testing data) per time-slot. It is given by:

$$\begin{aligned} &{s'}_{t+1} = mdp(s_t) \quad and \quad s'_{t+1} \neq s_{t+1} \\ &s_{t+1} \in set(sHD + randomLHD) \end{aligned} \tag{3}$$

Here, subscripts are time steps; s'_{t+1} is obtained via the function $mdp(s_t)$ which takes the previous state s_t as input. Thus, s'_{t+1} does not belong to the set of SHD or SLD states.

Reduction in such clashes indicates that MDP can learn user state preferences, even if they are costly in terms of power consumption. Now, clash rate reduction should preferably occur for these states, since we need to preserve what the user likes and let the MDP work on the remaining dispensable states which can help us to reduce the power consumption.

The first requirement shows that MDP understands the user preferences, and further, we can justify that our system achieves its primary objective of optimizing power consumption along with user comfort. We show that this power consumption efficiency comes from swapping most SLD states with the state outputs of the MDP.

IV. Experiments

All experiments have been performed on Ubuntu-16.04, Intel i5 Processor, and 16GB memory, which indicates that the proposed model can be deployed on personal computers efficiently.

A. Data-Set Description

For this study, we apply a real-life data-set that is part of the UMass Trace Repository and taken from Smart* Data-set for Sustainability[1]. Barker et. al. present a way to collect long-term data, which makes it easier and efficient to map a users behaviour [15]. The data contains the average electricity usage of several smart homes collected over regular intervals of 15 seconds for the following devices: Refrigerator, Freezer, Washer Dryer, Washing Machine, Dishwasher, Computer, Television, and Electric Heater. These devices have smart meters attached to them to record the power consumption after regular time intervals. We have chosen the data for smart homes A, B, and C. The data-set collected here is focused on sensing depth, i.e., collecting as much data as possible from each home, rather than breadth, i.e., collecting data from as many homes as possible.

Further, the data-set is divided into two parts: train and test. We use the training data for calculating MDP parameters, and the testing data to simulate live user interaction with our system. There is no distinction between training and testing data other than the fact that the training data is historical. Therefore, we assume it to be available from the beginning of the MDP run. Hence, all pre-processing and analysis are done on it. On the other hand, the testing data is present as live data with which we try to simulate the actual interaction of MDP with the user.

B. Domain State Clustering

TABLE I: Hyper-parameters for Growing Neural Gas Clustering.

GNG Parameters	Device Mode	Domain State
Max No. of Nodes	10000	20000
Max. Edge Age	100	50
Decay Rate: After Split	0.5	0.3
Decay Rate: Error	0.995	0.9

We considered clustering experiments using different algorithms such as DBSCAN, K-Means, and Growing Neural Gas (GNG). S. Dang explains the disadvantages of DBSCAN along with other clustering algorithms and states that density-based algorithms like DBSCAN do not take into account the topological structuring of the data, which is well mapped by the graphical modelling that GNG performs [16]. It is because of GNGs incremental nature that it is not necessary to decide on the number of nodes to use a priori, unlike the k-means clustering algorithm, where several trials may be required to determine an appropriate number of centres. GNG is, therefore, suitable for problems where we know nothing

[1]http://traces.cs.umass.edu/index.php/Smart/Smart

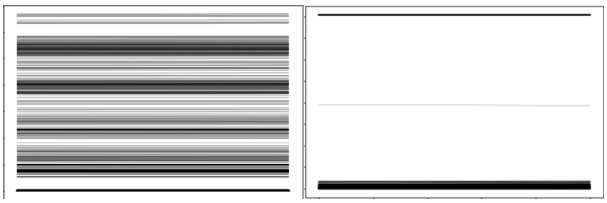


Fig. 3: Cluster Formations on the train data-set. Left: 263 domain state clusters obtained using GNG. Right: 3 device mode clusters obtained for Appliance 2 using GNG.

or little about the input distribution, or the cases in which deciding on network size and decaying parameters are very difficult or impossible. Hence, we base our results on GNG being the primary clustering algorithm.

Bernd explains the use of GNG in its entirety [17]. However, in our experiments, GNG had to be changed at certain steps where the algorithm is not explicit; for example, tackling the situation when a node does not have a neighbour and experimenting with the distance metrics used. We obtain device mode/state graphs and further find connected components that represent different data clusters. The algorithm is accurately able to find the number of classes and the associated neurons. Our prediction is based on the K-Nearest Neighbour approach upon GNG nodes, where we set k = 3. The hyper-parameters for clustering have been shown in I. The general parameters such as number of epochs is set to 150, the number of starting nodes is 1000, and the algorithm runs for 20 complete iterations before every single neuron is added.

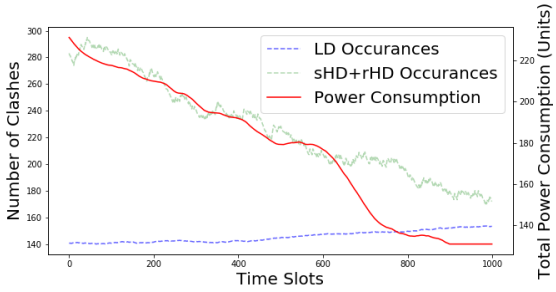
Other parameters are listed as follows:

1) 1) The value of top is found by studying the variance among the count of different individual domain states. We find that setting this to 22% yields the best results.
2) 2) We set fix to be 30% to simulate 1 in every 3 commands as strictly set by user regardless of MDP output. This has been set arbitrarily to enforce user authority.

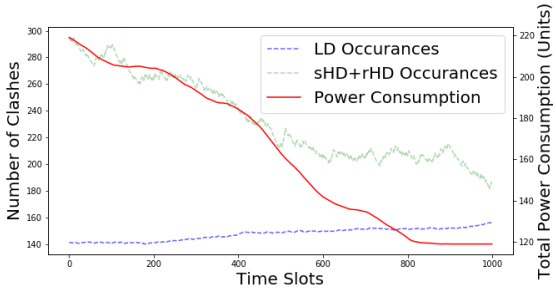
V. Results

We obtain 263, 242 and 240 cluster centres which make up the domain states, and, on an average, three modes per device, through Growing Neural Gas. We use 2,000,000 examples as our training data-se and. 1,000,000 examples as the test set. For validation, we divide the test set into 1,000 time slots, where each slot is essentially 1,000 consecutive readings. For each time slot, the clashes are calculated for SHD + SLD states. We also show LD state clashes along with Total Power Consumption averaged over each time slot. All the experiments for the data of each house have been conducted independently.

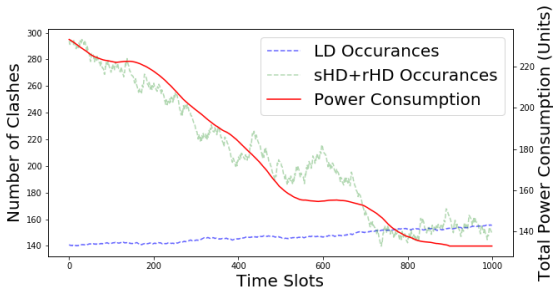
The results for clustering are shown in 3, which corresponds to graph outputs organized to show clusters for the GNG clustering algorithm. The left half of the image depicts 263 domain state clusters obtained in our experiments, and the right half of the image depicts three device mode clusters for one single appliance.



(a) for smart home A



(b) for smart home B



(c) for smart home C

Fig. 4: Resulting Clash Rates for LD, rLHD(randomLHD) Occurrences and corresponding fall in Total Power Consumption.

From Figure 4, we infer the following information :

1) Reduction in total Clash Rate, since clashes for LD states, is more or less constant, and clashes in SHD + SLD occurrences reduce over time along with the total clashes, confirming that MDP is learning user behavior by outputting a greater number of states which the user would “like” or prefer.
2) MDP is not only learning user behavior but is also being specific about learning SHD states. Thus, the system chooses not to alter the usual user requirements.
3) When the MDP is beginning to learn the behavior, it is natural to have high number of clashes. However, over time when the MDP is intelligent enough to predict states, it outputs a better state than the current LD state, thereby creating the required scope for improvement regarding power consumption.

TABLE II: Comparison between different methods.

Method	**Energy Saved (%)**	**Considers User Behaviour**
Autonomous Appliance Scheduling for Household Energy Management [7]	10.92	No
Intelligent Household LED Lighting System Considering Energy Efficiency and User Satisfaction [11]	21.9	Yes
Making Smart Home Smarter: Optimizing Energy Consumption with Human in the Loop	30	Yes

4) Total Power Consumption (in units) reduces over time, and this reduction is in a nearly linear fashion, in line with the decline in clash rate for SHD + SLD occurrences. The average power consumption is around 200 points on power consumption scale without the intelligent planner. In all the cases, we achieve convergence before 800 time-slots. This inference is logically consistent with our observation that the MDP is giving the user enough time to exhibit their usage habits and hence, adapts accordingly.
   In figure 4, an exciting behavior is observed between the LD and SHD + SLD occurrences in the plots for the houses. We elaborate on the same as follows:

5) Around the time slot just before convergence, the plots in figure 4 contrast to our theory that with a reduction in SHD + SLD clash occurrences, LD state clashes would compensate and increase. Figure 4c provides evidence at time slot 760 in which LD states increase linearly, whereas there are abrupt changes in SHD + SLD occurrences. This behavior, however, further bolsters the point that our model is adjusting to determine precisely which LD states need to be ignored and which SHD + SLD states are to be preferred.

The results show that considering user behavior while trying to reduce power consumption is a significant factor because the user must agree with the system when it comes to operating devices around him. In Table II, we show how the adaptive configuration of smart devices is quantitatively better than the other existing techniques. The mentioned methods are the closest to our problem statement, however, not exact. As explained in Section II, Byun et. al. do not perceive user behavior to the extent as our method does, while Adika et. al. do not take user preference into account at all [7], [11]. Their result shows around 22% energy saving as opposed to a much better 30% energy saving achieved by our method. Hence, our work justifies which states are compromised to obtain energy efficiency, and performs better at this task while considering user device setting preferences against other methods.

VI. Conclusion

This paper presents an approach to adaptively configure the smart IoT devices through the effective incorporation of probabilistic user behavior(s). It takes into account the variance of devices regarding their application and more importantly, the user preference. However, it looks at the data only through power consumption perspective. Simulation of user actuation is intentionally kept stochastic to showcase that our system is resilient enough for real-life scenarios. The system uses each type of information available in the data-set, albeit with certain assumptions which at no point violate the requirement of stochasticity. Moreover, the system is made robust via the use of multilevel clustering. LD/HD states and the MDP algorithm help maintain an effective trade-off with user-preferred states and states which consume less power.

Even though current results very well substantiate the effectiveness of our system, we intend to extend our experiments to accommodate environments where even the clustered domain states are exponential. This problem can be tackled via enhanced state space approximation techniques and reduced computational requirements using factored MDPs. Future work can include formulating novel reward functions, which are a function of time as well, unlike the current static function since reward functions may be a function of time or the frequency of the number of shifts between device modes, etc. Also, we intend to search for a way to cluster devices based on their application in an automated manner to strengthen our automatic intelligent planner further.

References

[1] Yoram Chisik. An image of electricity: Towards an understanding of how people perceive electricity. In Pedro Campos, Nicholas Graham, Joaquim Jorge, Nuno Nunes, Philippe Palanque, and Marco Winckler, editors, *Human-Computer Interaction – INTERACT 2011*, pages 100–117, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[2] Willett Kempton and Laura Montgomery. Folk quantification of energy. *Energy*, 7(10):817 – 827, 1982.

[3] Melanie R. Herrmann, Duncan P. Brumby, Tadj Oreszczyn, and Xavier M. P. Gilbert. Does data visualization affect users understanding of electricity consumption? *Building Research & Information*, 46(3):238–250, 2018.

[4] Murray Goulden, Ben Bedwell, Stefan Rennick-Egglestone, Tom Rodden, and Alexa Spence. Smart grids, smart users? the role of the user in demand side management. *Energy Research & Social Science*, 2:21 – 29, 2014.

[5] Jenny Palm, Kajsa Ellegrd, and Mattias Hellgren. A cluster analysis of energy-consuming activities in everyday life. *Building Research & Information*, 46(1):99–113, 2018.

[6] S. S. Yau and A. B. Buduru. Intelligent planning for developing mobile iot applications using cloud systems. In *2014 IEEE International Conference on Mobile Services*, pages 55–62, June 2014.

[7] C. O. Adika and L. Wang. Autonomous appliance scheduling for household energy management. *IEEE Transactions on Smart Grid*, 5(2):673–682, March 2014.

[8] M. Jahn, M. Jentsch, C. R. Prause, F. Pramudianto, A. Al-Akkad, and R. Reiners. The energy aware smart home. In *2010 5th International Conference on Future Information Technology*, pages 1–8, May 2010.

[9] Markus Eisenhauer, Peter Rosengren, and Pablo Antolin. Hydra: A development platform for integrating wireless devices and sensors into ambient intelligence systems. In Daniel Giusto, Antonio Iera, Giacomo Morabito, and Luigi Atzori, editors, *The Internet of Things*, pages 367–373, New York, NY, 2010. Springer New York.

[10] C. Wei and Y. Li. Design of energy consumption monitoring and energy-saving management system of intelligent building based on the internet of things. In *2011 International Conference on Electronics, Communications and Control (ICECC)*, pages 3650–3652, Sep. 2011.

[11] J. Byun, I. Hong, B. Lee, and S. Park. Intelligent household led lighting system considering energy efficiency and user satisfaction. *IEEE Transactions on Consumer Electronics*, 59(1):70–76, February 2013.

[12] Jose Costa and Ricardo Oliveira. Cluster analysis using growing neural gas and graph partitioning. pages 3051 – 3056, 09 2007.
[13] J. R. Norris. *Markov Chains*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1997.
[14] Bernhard Nebel Wolfram Burgard and Martin Riedmiller. Acting under Uncertainty: Choosing Individual Actions, Markov Decision Processes (MDPs), Value Iteration. http://gki.informatik.uni-freiburg.de/teaching/ss14/gki/lectures/ai13.pdf, 2014.
[15] Sean Barker, Aditya Mishra, David Irwin, Emmanuel Cecchet, Prashant Shenoy, and Jeannie Albrecht. Smart*: An open data set and tools for enabling research in sustainable homes.
[16] Shilpa Dang. Performance evaluation of clustering algorithm using different datasets. *IJARCSMS*, 3:167–173, 01 2015.
[17] Bernd Fritzke. A growing neural gas network learns topologies. In *Proceedings of the 7th International Conference on Neural Information Processing Systems*, NIPS'94, pages 625–632, Cambridge, MA, USA, 1994. MIT Press.

Modelling Of Viscoelastic Fluid In Finishing Processes

Krovvidi Srinivas, Q. Murtaza, A.K Agarwal

Abstract: The Traditional Finishing process mostly lapping, honing, buffing, superfinishing etc contains various disadvantages like labour intensive work, high Production cost and time. Moreover, the quality of surface finish achieved is not up to the mark but in present era there is requirement of better quality of products with desired surface finish, in order to overcome this finishing issue Abrasive Magnetic Finishing and Abrasive Flow Machining process comes into the consideration. Both the Process uses a Viscoelastic media which contains an adequate amount of Polymer, gel, Oil and abrasives. Magnetic Abrasive Finishing includes polymer media with carbonyl particle which align along the magnetic field and results in effective machining. This paper includes solid modelling of the process and simulation on Ansys Fluent module and analyse the Flow parameters of the viscoelastic polymer media and its effect on material removal.

Index Terms: viscoelastic Magnetic Abrasive Finishing (VMAF), Magnetic Abrasive Finishing (MAF), Material Removal (MR), Surface Finish, Ansys Maxwell, Polymer Media, Abrasive Flow Finishing (AFF).

1 INTRODUCTION

Magnetic abrasive finishing is an effective machining process in which magnetic field is used to machine the target surfaces of the workpiece by influencing the carbonyl particles and directing them towards the wall of the workpiece for machining. In conventional abrasive machining process both the workpiece and the target surface are polarised and due to the interplay of magnetic field, a Brush of abrasive is formed which results in the effective machining of the workpiece. This effective concept of machining got the eyeball of many scholar and initiated a large number of researches in this field like Kang and Yamaguchi [1] developed a multipole dip- system for the machining of internal surface of the capillary tube. The researchers highlighted the difficulties faced in machining of the capillary tube mainly in controlling the magnetics abrasive in MAF process, the researchers suggested that the multipole system with alternate magnetics and nonmagnetic region simultaneously applied on the workpiece which results in the effective machining of the long capillary tube. This process has allowed the application of MAF on long Capillary tube as well which was earlier confined to the short length tube only. The final conclusion of the researcher was that the double pole system finishes the workpiece in half a time as compare to the single pole system. V.K Jain [2] explained various non-traditional machining process which are used in modern times which involves the use of abrasive and the media to support it. The various process which he discussed in its paper was Abrasive Flow Finishing (AFF), Magnetic Abrasive Finishing (MAF), Magnetorheological Finishing, Magnetorheological Abrasive Flow Finishing, Elastic Emission Machining (EEM) and Magnetic Float Polishing. he laid attention on the magnetic abrasive finishing process and gave some important conclusion like involvement of slot in the magnet over non machining zone and the use of pulse Dc supply rather than smooth DC supply in the system. Wani et al. [4] gave stress on Magnetic Abrasive flow finishing which is a process made by combining the advantage of both abrasive flow machining and the magnetic abrasive finishing process, and used finite element method to find distribution of magnetic potential in the magnetic brush. Through this distribution the researcher evaluated machining pressure, surface finishing and material removal. After the successful evaluation of the results the researcher claims that these results are similar to those present in the literature. Kala et al. [4] developed a setup which aims to catch the force signature for double disc magnetic abrasive finishing process and the conventional abrasive magnetic abrasive finishing process and by these properties the researchers aims to understand the working of the flexible magnetic abrasive brush. Thus, to do this the researchers performed the magnetic abrasive finishing operation in single as well as double disc and finally understood the property of flexible magnetic abrasive brush behaviour from it. Wu et al [5] developed a model of magnetic abrasive finishing process and analyzed its properties based on the rotational speed, cutting fluid and the magnetic pole and current frequency and its effect on change in material removal and surface finish are invested. The researchers finally claimed that the neat cutting oil is required for processing which results in higher material removal and finer finish. They also concluded that finishing is better with the increasing in rotational speed and as the current frequency increases the angle variation of the particle increases and if frequency is low the surface requirement is up to few nanometers. Rathod et al. [6] correlated the difference shape factor with wide range of particle size. The researcher made use of slurry abrasion tester with silica sand slurry and commented on the effect of sliding distance slurry abrasion volume, where they argued that volume loss is proportional to the sliding distance. They also studied the effect of micro ploughing and cutting. Srivastav et al. [7] simulated a model of electrochemical magnetic abrasive finishing process in which the researchers used stainless steel workpiece whose radius was ten mm and thickness was 2.5 mm. the researchers used chromium oxide as the abrasive particle whose average diameter was 387 micrometers, the intensity of magnetic field was around 1 tesla and the rotational speed of tool was taken as 5835 rpm and its diameter was 6 mm. the electrolyte used for the experiment was brine solution with twenty percent concentration. The researchers study the thermal model in both steady and transient model. Srivastav and Amit [8] did the thermal analysis of work-brush interface and along the workpiece depth in magnetic abrasive finishing process with the help of ANSYs R15 software in both steady and thermal state. The value of magnetic field intensity given to 6 mm diameter tool was 1 tesla. The researcher used chromium oxide abrasives of about 387 micrometer Diameter. The workpiece has a radius of 10 mm and 2.5 mm thickness and the material of the workpiece was silicon nitride. Singh et al. [9] used Taguchi L orthogonal array to design an experiment which could measure surface temperature of the finishing surface of mild steel using MAF process. The researchers used current, working gap, Rotational speed and abrasive weight as

its important process parameters. After the Taguachi optimization the Fuzzy logic is used by researcher through which prediction of surface temperature is and also it is validated. Jaiswal et al. [10] made FEM model of MAF and investigated the distribution of magnetic force on the workpiece. The researcher also made a theoretical model for computing the material removal and surface roughness. For this the workpiece was considered of uniform surface without statically distribution. Numerical experiments were also computed which were validated later on through literature. Yan et all [11] performed the experiment on MA genetic abrasive finishing based on a simulated model which is based on FEM model of electromagnetic inductor. He then compared the simulated model with the actual electromagnetic inductor and predicted a marginal error of only 7 percent. Li et all [12] developed a finishing setup having new media for finishing the rotatory surface ANSYS Maxwell was used to determine the optimum angle between the poles and the mathematical model was made to compute the MRR in which pressure and velocity was computed based on arced wear model. The MRR computation model was made which predicted MRR as a function of pole rotational speed, Magnetic flux density, cam rotational speed and the ferromagnetic and abrasive particles' diameter. Li et all [13] proposed a new viscoelastic magnetic Abrasive based on analysis of field characteristic and existing finishing process. They also discussed the motion locus of the abrasives. The present scenario, abrasive flowing finishing, grinding, abrasive brush deburring, magnetorheological finishing, manual deburring etc. are broad name in term of finishing but still they possess some limitations in finishing quality and efficiency. Shimamura et al. [14], [15] modelled and tested that magnetic abrasive particles are subjected to the pressure for machining and that pressure is a function of magnetic flux density, number of abrasive particles and the permeability of abrasive medium. Fox et al [16] observed that the magnetic force (function of the volume and magnetic susceptibility of the ferrous particles in the magnetic field, magnetic field intensity and the gradients at the finishing area). Srinivas et al. [17] studied the distribution of the magnetic field intensity along with the computation of torque and force and related flow parameters of Viscoelastic Magnetic Abrasive finishing process with magnetic field intensity. The Researchers used ANSYS Maxwell and ANSYS Fluent in order to find co relation between the flow parameters and magnetic field intensity and commented on the successful application of the proposed viscoelastic media. The researchers also plotted the graph between the magnetic field intensity and the current which is applied on the single pole electromagnet to produce the desired magnetic field for machining in the magnetic abrasive finishing process. Srinivas and Anant [18] studied the pressure variation in abrasive flow machining process with the help of a simulated model made on Ansys Fluent and commented on the effect of pressure variation in the machining operation. They used viscous media similar to that used in viscoelastic magnetic abrasive finishing with the difference that in later process there is an addition of carbonyl particles which gets activated by magnetic field. The researcher computed pressure difference between 40 MPa and 20 MPa and the value of pressure so obtained was validated with the help of literature review. Ali et al. [19] reviewed various researches on AFM which uses a viscoelastic media and semisolid media which can support various abrasive based finishing process including both magnetic and non-magnetic based finishing process. Bhardwaj et al [20] compiled the researches which contained the use of viscoelastic media in various hybrids of abrasive flow machining process, the researchers also commented on the tooling and fixturing in various hybrids of AFM and the compatibility of the viscoelastic media with the hybrid process. So it can be seen that Viscoelastic Magnetic Abrasive Finishing has a vital role in Finishing department and must be consider for effective machining operation.

2 SIMULATION AND MODELING

A simulation model was proposed to overcome the research gap in the process of magnetic abrasive finishing process. For this purpose, ANSYS Fluent was used so a basic geometry was made in CREO 3.0 software and was then transferred to ANSYS fluent as an IGES File. The basic model contains a brass workpiece with alumina abrasive mixed in the viscous fluid whose property are listed in the tables. When a geometry is port into the ANSYS fluent software, it undergoes 5 steps. The first step (Figure 1) is the geometry where we do name selection and perform the Boolean operations if required. So, we did name selection in which the selections were Inlet, Outlet, workpiece and fluid Domain, next was the Boolean operation where we created path for the fluid flow within the internal surface of the workpiece. Then the file was transferred into the meshing section in which the mesh was created which is depicted by figure 2 and 3. The total elements in mesh was about $3X\ 10^3$ and the shape of the element was majorly tetragonal. After successful meshing the next step was to load the fluent setup for which is shown in figure 4. This is the major step in which we select the type of simulation, energy equation, material assignment, boundary condition and select the number of iterations for the process of simulation. For the process the laminar simulation with the boundary condition on was chosen and the adequate material is assigned to the parts. The boundary condition used for the simulation was Pressure inlet of 40 MPa and the outlet at 20 MPa, the workpiece and walls were stationary. The detail of boundary condition is given in table 2. Initially the fluid is not given any velocity but its movement would be there through the pressure difference which would do machining. The effective machining may be done when we will be placed magnet inside the system which cannot be done on ANSYS fluent, although for that ANSYS Maxwell must be taken into consideration. After all this assignment the simulation is allowed to run for iteration and the plot of energy along with the X, Y and Z velocity was plotted against the number of iterations, which is shown in figure 6.

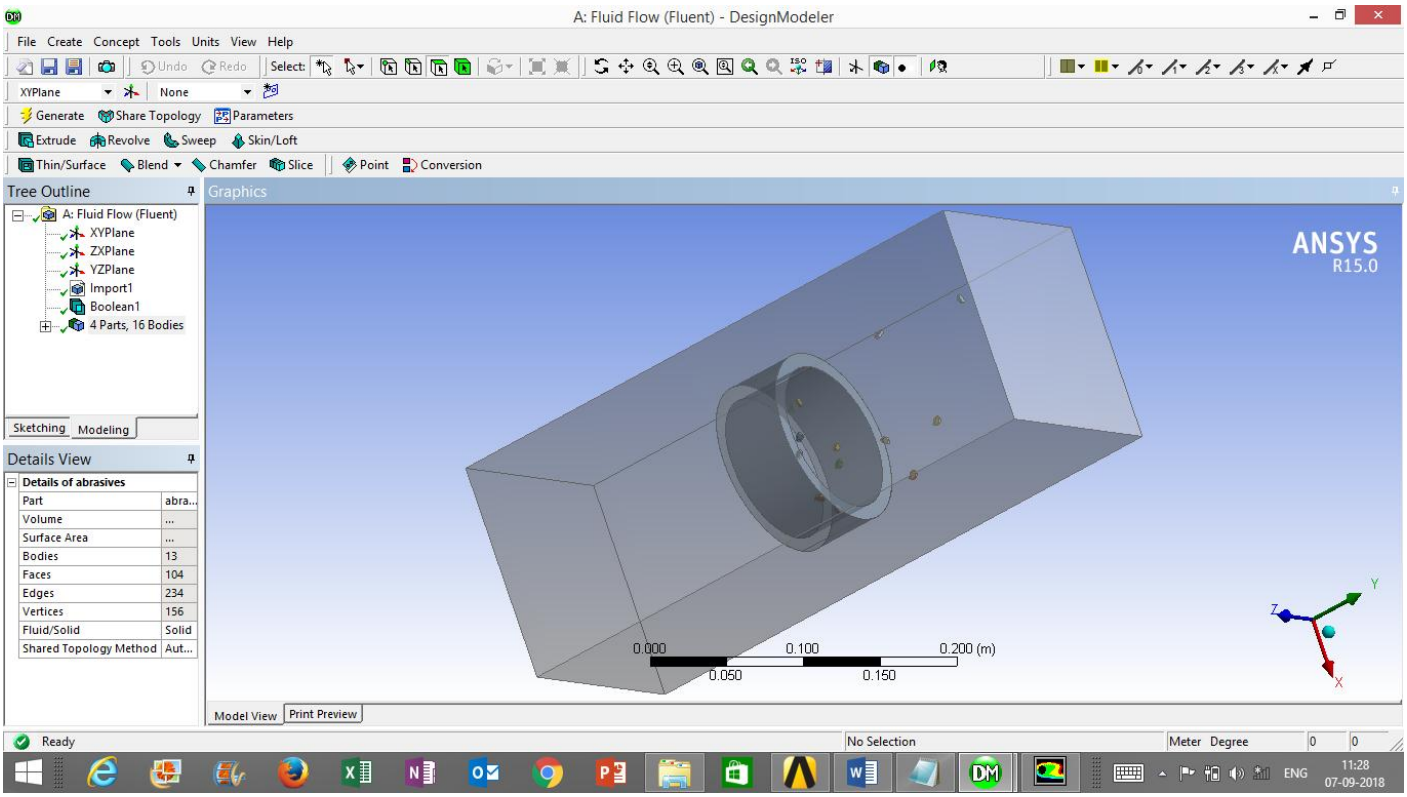


Fig.1. *Initial geometry in ANSYS fluent*

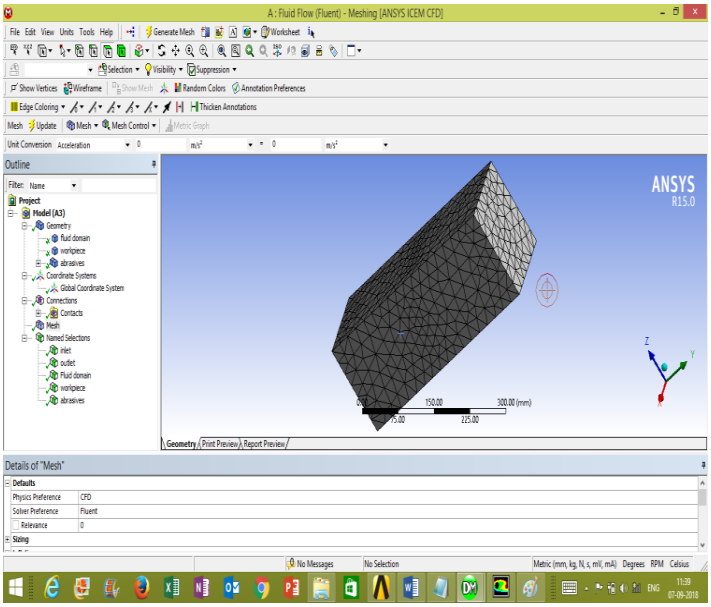


Fig. 2. *Meshing of Fluid Domain*

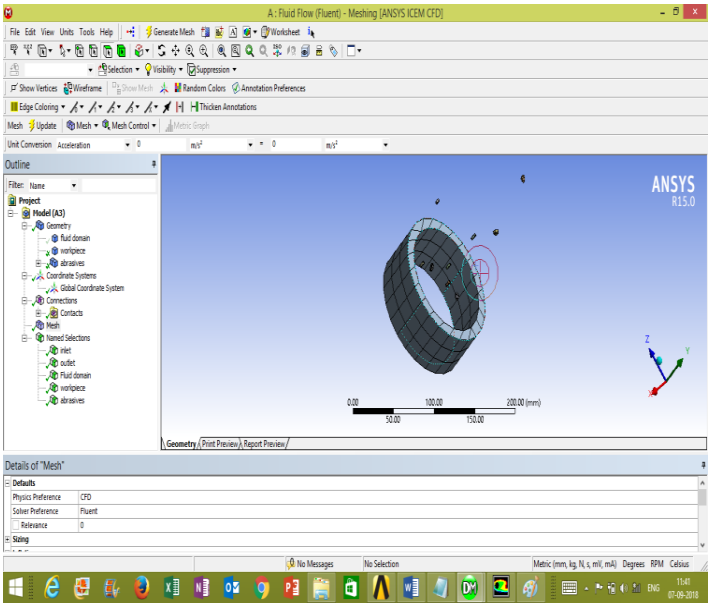


Fig. 3. *Meshing of work piece*

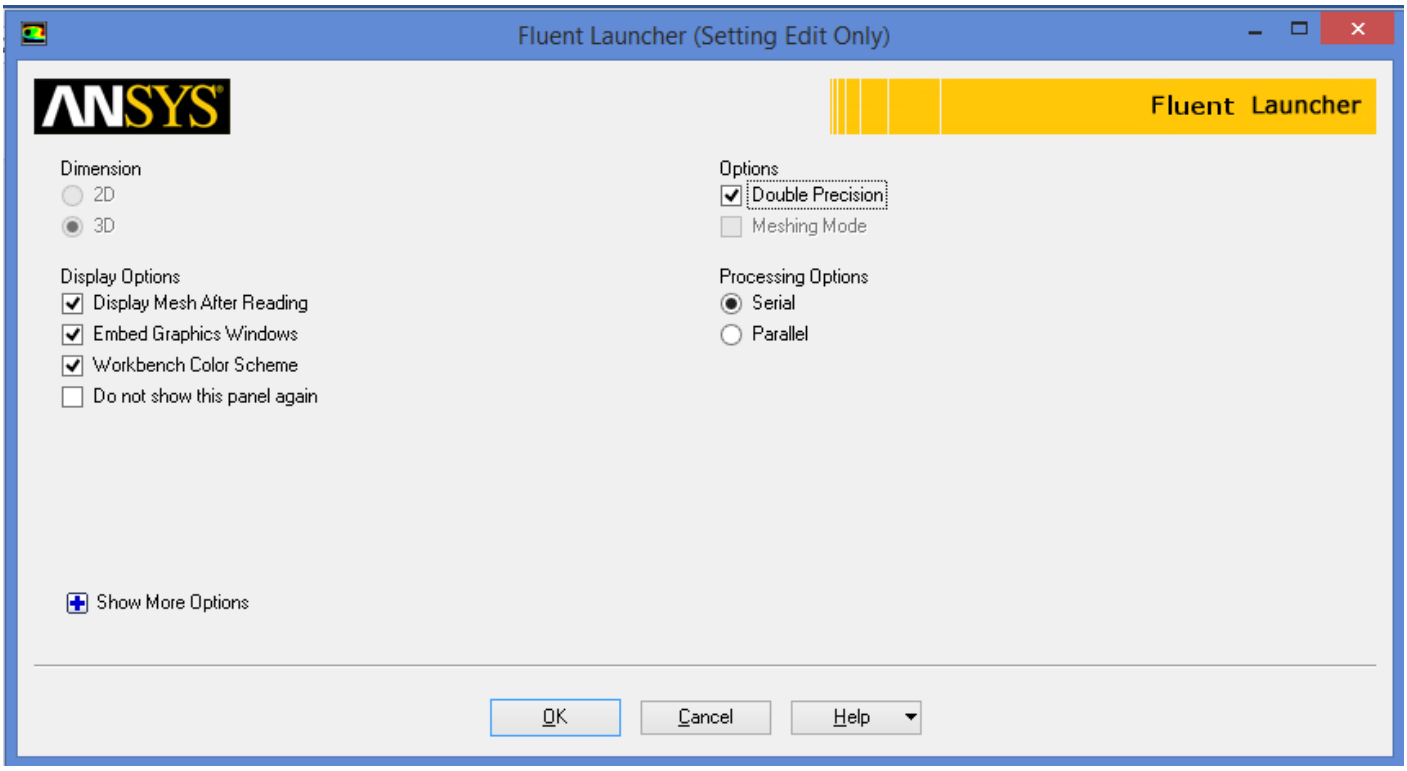


***Fig.4**. Initial setup*

TABLES

***Table 1** Showing the Domain Nodes and Elements of Simulations*

MATERIAL	TYPE	DENSITY (Kg/m³)	VISCOSITY (Kg/m-s)	THERMAL CONDUCTIVITY (W/ m.K)	SPECIFIC HEAT (J/Kg-K)
Viscous Fluid	Fluid	1219	0.739	0.22	2025
Brass	solid	8500		109	400
Aluminum abrasives	solid	3950		12	451

***Table 2** showing The Boundary condition of simulations*

Sr no	Name Selection	Boundary condition
1.	Inlet	Inlet pressure: 40 MPa
2.	Fluid	Stationary fluid at room temperature: 300 K
3.	Workpiece	Stationary workpiece at Room Temperature: 300 K
4.	Outlet	Outlet Pressure: 20 MPa

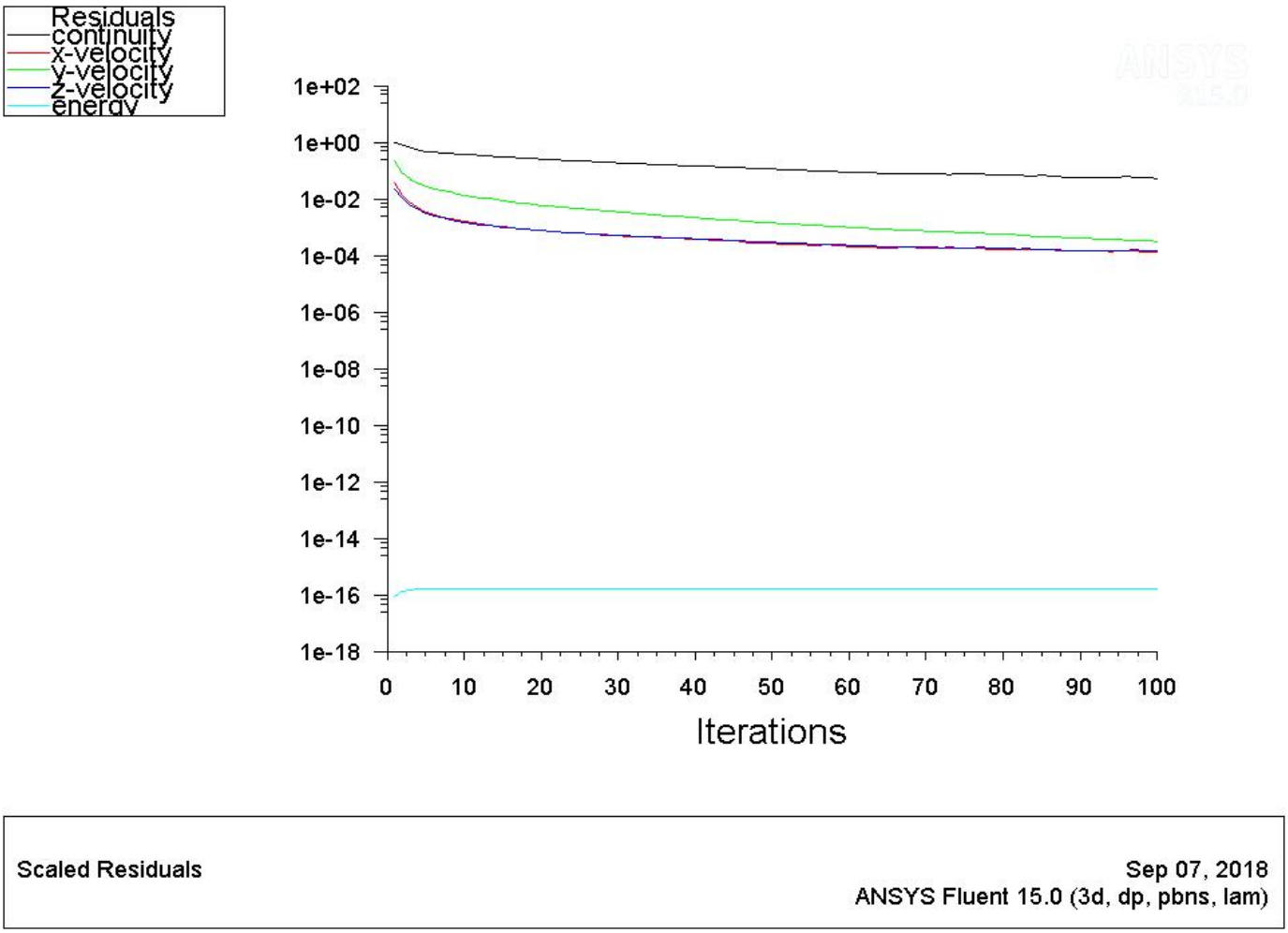


***Fig**. 5. Energy Vs Iteration curve*

3 Results and Discussions

As The results show the maximum pressure is 26.2455 MPa and the minimum Pressure obtained after the analysis is 10.4519 MPa which is an optimum value for machining the brass workpiece at the pressure range of 40 MPa to 20 MPa (Figure 7). The pressure drops in the reason of the workpiece due to the bell effect, When the surface area decreases at the neck region the velocity of the media decreases resulting in the increase in the pressure value at the neck region which is at the starting of workpiece in the direction of the flow, But the value of the pressure decreases throughout the flow of the polymer media which can be seen through the subsequent figure in which the pressure variation along the plane is given, in which the drop in the pressure is shown. Figure 8 shows the streamline variation of the Polymer media. As observed from the figure the stream lines are uniform and smooth and representing the flow due to the pressure difference only. The effective machining may also be achieved by putting a magnet around the workpiece which will attract the carbonyl particle present in the media which would further hit the magnetic particles. The figure 9 shows the force applied by the polymer media on the workpiece, though the force applied would be uniform as the abrasive would be uniformly mixed with the polymer media which results in the application of force on the workpiece. Though the maximum value of the force is 5440 N. The force would be maximum at the point where the abrasives strikes the surface of the workpiece. The figure 10 shows the distribution of velocity of the polymer media on the workpiece, the polymer media is specially made by combining various polymer, gel, oil and abrasives blended together to form a semisolid putty which is extruded up and down with the pressure difference to achieve the effective machining of the process. As Described by the process the maximum velocity is 183.2 m/s and the minimum velocity is 6.94 m/s, again this point must be taken care of that the figure is showing the velocity distribution at the region of the workpiece and not along the flow. The distribution of velocity along the flow would be different. The figure 11 shows the pressure variation along the entire length of the cylinder in the abrasive flow machine rather than the workpiece regions. It shows the pressure at the upper side is 29 MPa and at the lower side is 10.45 MPa the region for the pressure drop and is already discussed. Hence simulation was successful in showing machining characteristic of Viscoelastic media

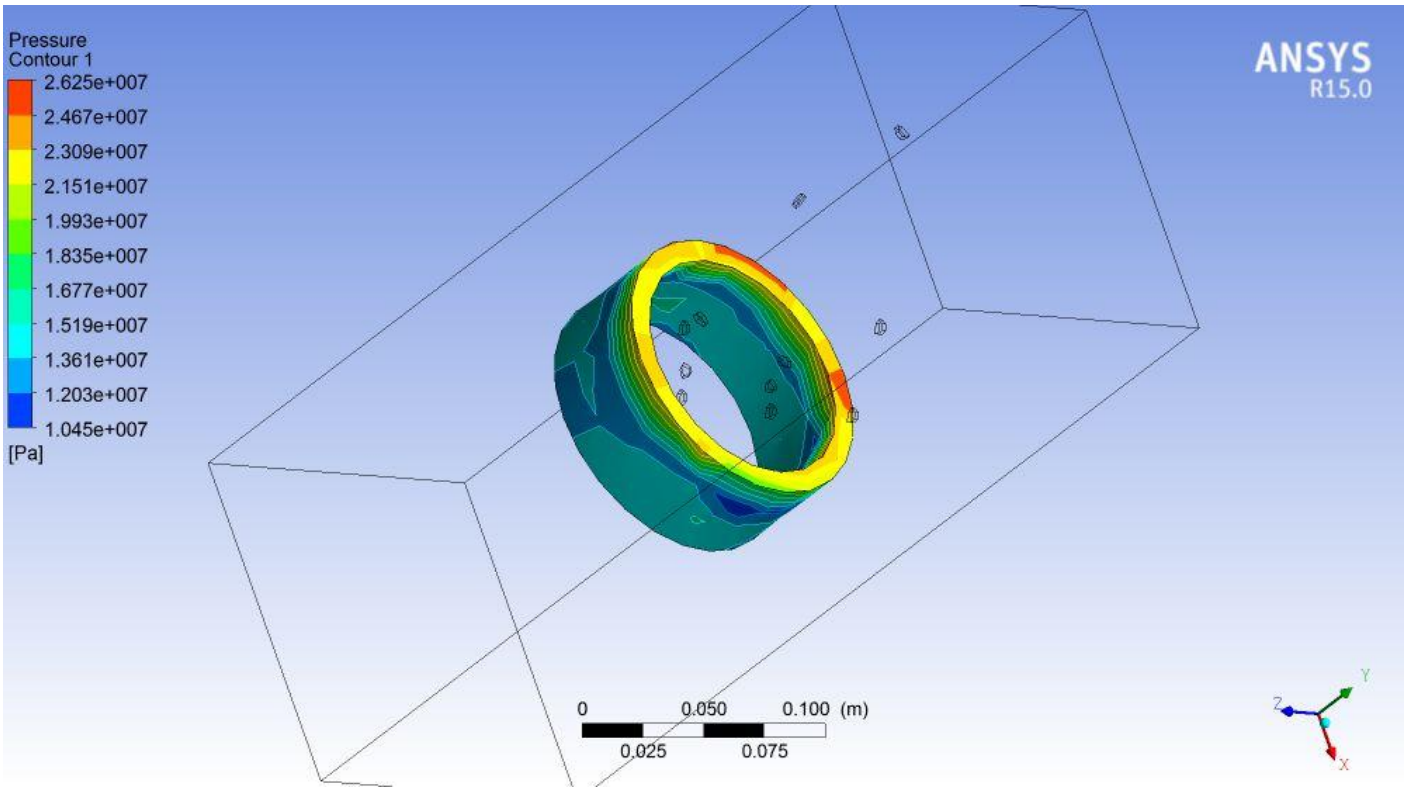


FIG.6 *EFFECT OF PRESSURE VARIATION ON WORKPIECE*

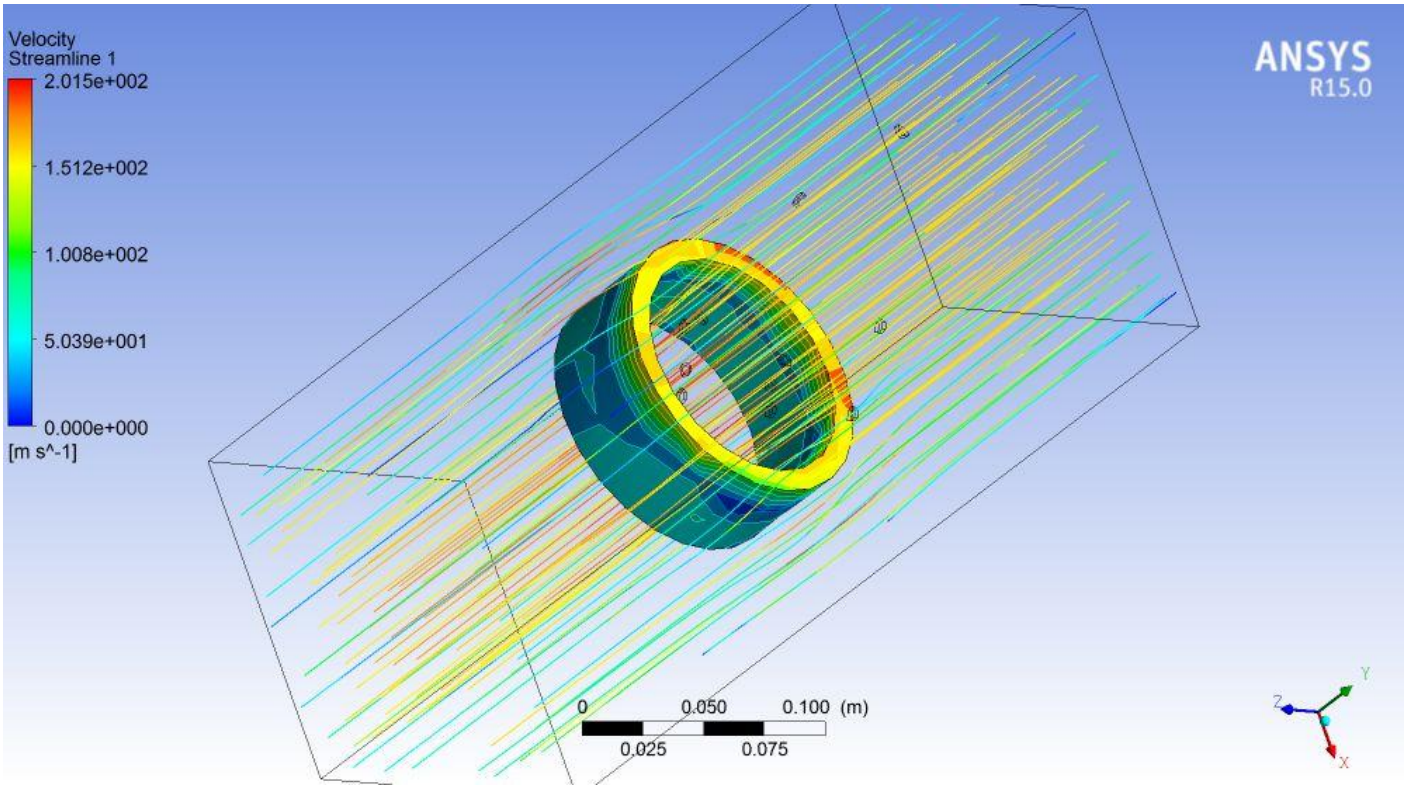


FIG.7 *STREAMLINE VARIATION OF POLYMER MEDIA*

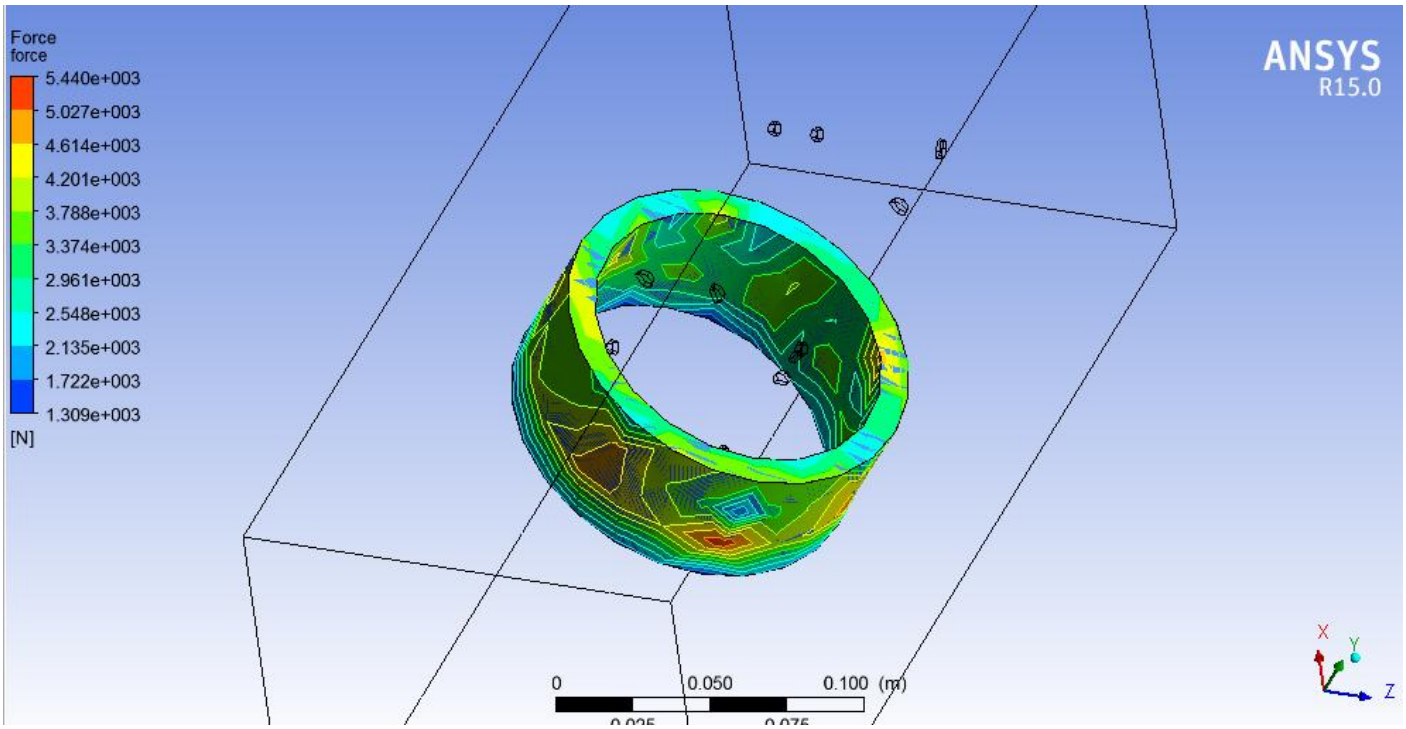


FIG.8 *FORCE APPLIED ON POLYMER MEDIA ON WORKPIECE*

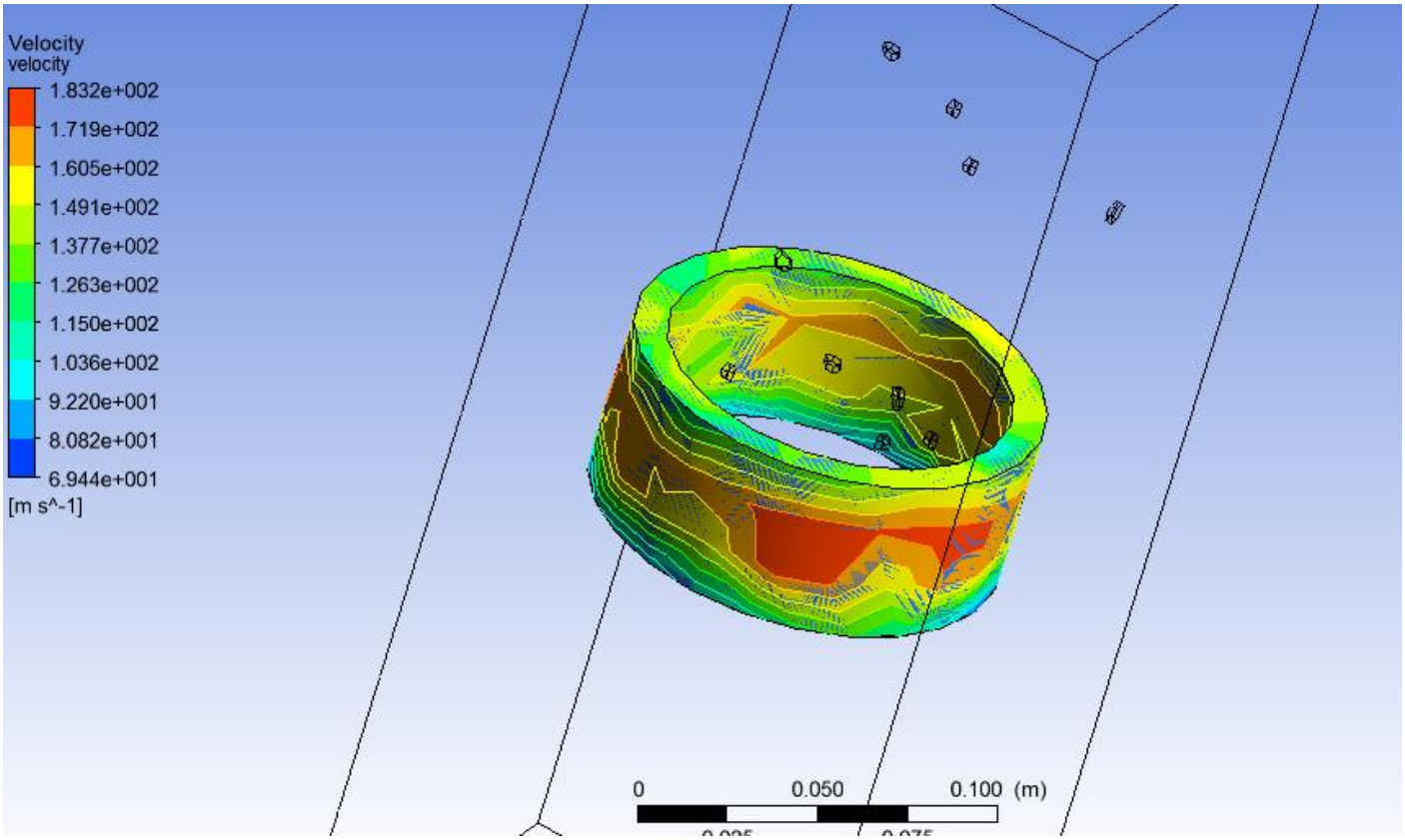


Fig 9 *VELOCITY OF POLYMER MEDIA*

IJSTR©2019
www.ijstr.org

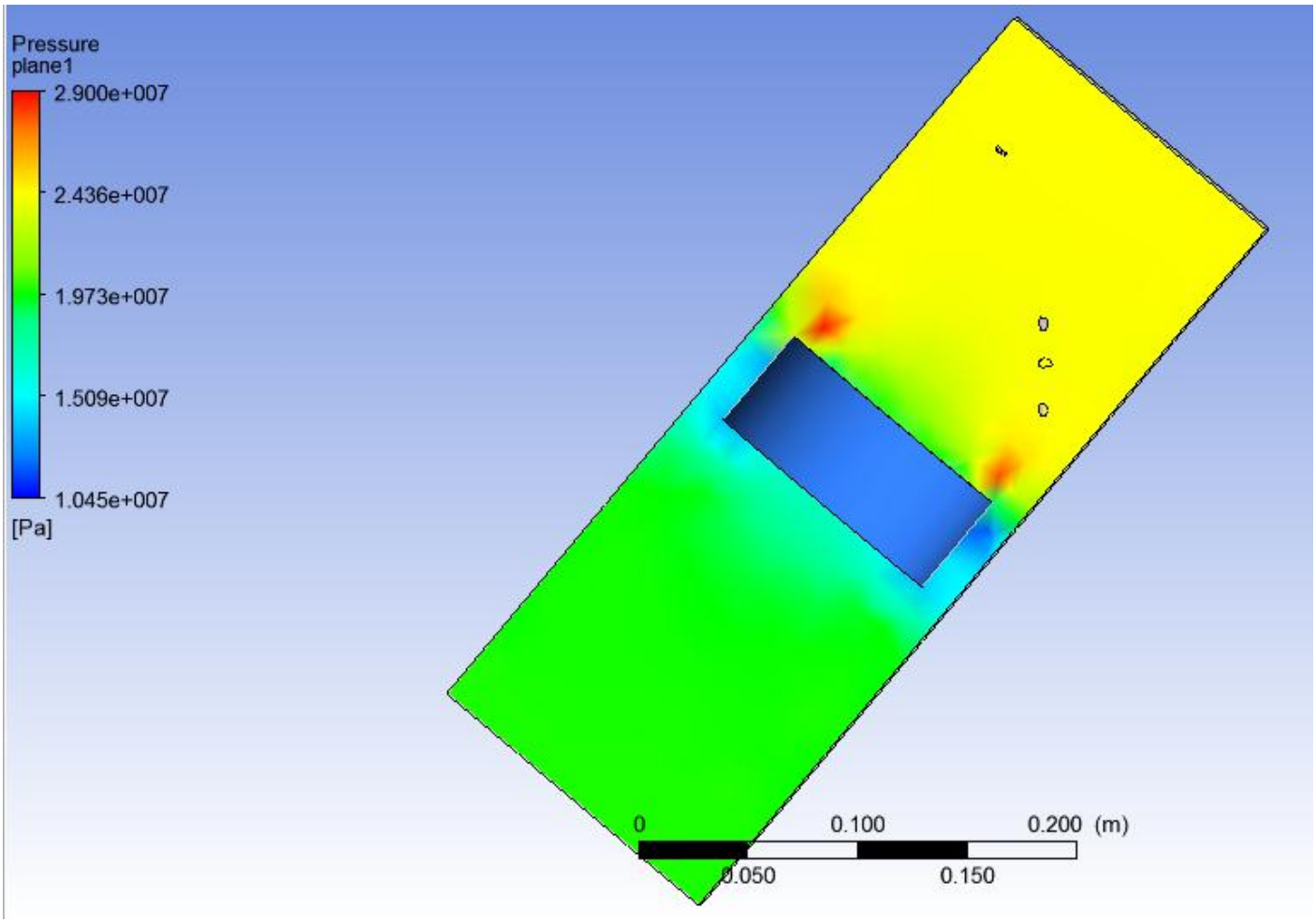


Fig 10 *PRESSURE VARIATION ALONG POLYMER MEDIA*

REFERENCES

[1] Kang, Junmo., Yamaguchi, Hitomi: Internal finishing of capillary tubes by magnetic abrasive finishing using a multiple pole-tip system, Precision Engineering 36 (2012) 510– 516, 2012.

[2] Jain, V. K: Magnetic field assisted abrasive based micro-/nano- finishing, Journal of Materials Processing Technology 209 (2009) 6022–6038, 2009.

[3] Wani, Amit. M., Yadava, Vinod., Khatri, Atul: Simulation for the prediction of surface roughness in magnetic abrasive flow finishing (MAFF), Journal of Materials Processing Technology 190 (2007) 282–290.

[4] Kala, Prateek., Pandey, Pulak. M., Verma, Girish. C., Sharma, Varun: Understanding flexible abrasive brush behaviour for double disk magnetic abrasive finishing based on force signature, Journal of Manufacturing Processes 28 (2017) 442–448, 2017.

[5] Wu Jinzhong., Zou, Yanhua., Sugiyama, Hitoshi: Study on finishing characteristics of magnetic abrasive finishing process using low-frequency alternating magnetic field, Int J Adv Manufacturing Technology 85:585– 594 DOI 10.1007/s00170-015-7962-9, (2016).

[6] Rathod, Avishkar., Sapate, S. G., Khatirkar, R. K: Shape factor analysis of abrasive particles used in slurry abrasion testing, International Journal of Mechanical and Industrial Engineering (IJMIE) ISSN No. 2231 –6477, Vol-2, Iss-4, 2012.

[7] Srivastava, Ankit. Kumar., komma, V.R: Finite Element Modelling of Electrochemical Magnetic Abrasive Finishing, Imperial Journal of Interdisciplinary Research (IJIR), Vol- 2, Issue-12, 2016, ISSN: 2454-1362.

[8] Srivastava, Ankit. Kumar., Katiyar. Amit.: Modelling and Simulation of Magnetic Abrasive Finishing for Thermal Analysis, Imperial Journal of Interdisciplinary Research (IJIR) Vol-2, Issue-12, 2016 ISSN: 2454-1362.

[9] Mr. Krovvidi Srinivas is presently working as Assistant Professor in Delhi Technological University. He is having a total teaching experience of 24 years. He guided around 75 UG projects and 7 PG projects. He published around 10 research papers in different conferences and Journals.He is a Member of Institution of Engineers (I), Life member of Indian Society for Technical Education, Life member of Powder Metallurgy Association of India.

[10] Presently, Dr. Qasim Murtaza is a Professor in the Department of Mechanical Engineering, Delhi Technological University, Delhi, India. He got his Ph.D (2006) in the area of manufacturing process, funded

by European Union as Research Assistant, from Metal Thermal Spray Lab, Dublin City University, Ireland. He has also published several papers in reputed SCI journals and conference proceedings.

[11] Dr. A K Agrawal is presently working as Professor in the Mechanical Engineering Department, Delhi Technological University. He is Having more than 30 years of Teaching and research experience. He published several papers in International journals and Conferences.

[12] Singh, Rajneesh. Kumar., Singh, D.K., Gangwar. Swati.: Fuzzy logicoptimization of process parameters affecting surface temperature in MAF, IOP Conf. Series: Materials Science and Engineering 404 (2018) 012017 doi:10.1088/1757-899X/404/1/012017.

[13] Jaiswal, S.C., Jain, V. K., Dixit, P.M.: Modelling and simulation ofmagnetic abrasive finishing process, Int J Adv Manufacturing Technology (2005) 26: 477–490, DOI 10.1007/s00170-004-2180.

[14] Yan, Q. S., Qiu, T.X., Gao, W.Q.,Meng, L: Simulation and Experimental Analysis of Electromagnetic Inductor for Magnetic Abrasive Finishing, Key Engineering Materials Vols. 389-390 (2009) pp 193-198 Online: 2008-09-26 © (2009) Trans Tech Publications, Switzerland, doi: 10.4028/www.scientific.net/KEM.389- 390.193.

[15] Li, Wenhui., Li Xiuhong., Yang, Shengqiang., Li, Weidong: A newly developed media for magenetic abrasive finishing process: Material removal behaviour and finishing performance, Journal of Materials Processing Tech 260(2018) 20-29.

[16] 13.Li, Xiuhong., Li, Wenhui., Yang, Shengqiang.: Effect of coupling Agent on Interfecial Bonding Properties of Visecoelastic Magnetic Abrasive Tools and Finishing Performance, ISSN 1473-804 online, 1473-8031 print, DOI 10.5013/IJSSST.a.17.28.17.

[17] 14.Shinmura T, Takazawa K, Hatano E, Matsunaga M. (1990) Study on magnetic abrasive finishing. Ann CIRP, 39(l): 325-328.

[18] 15.Shinmura T., Takazawa K., Hatano E., Aizawa T. (1985) Study on magnetic abrasive process-process principles and finishing possibility, Bull. Japan Soc. Prec. Eng. 19(1) 54–55.

[19] 16.Kim JD., Choi MS. (1995) Simulation for the prediction of surface accuracy in magnetic abrasive machining. J Mater Process Tech 53:630–642.

[20] 17.Srinivas. K., Murtaza. Q., Aggarwal. A.K.: Variation of Flow parameters and Magnetic Flux in Viscoelastic Magnetic Abrasive Finishing process, ISSN 0973-4562 Volume 14, Number 8, pp. 1940-1946, (2019).

[21] 18.Srinivas. K., Bhardwaj, Anant.: Pressure variation in Abrasive Flow machining: Modelling and simulation, 3rd international conference on Advanced Production and Industrial Engineering (ICAPIE'18), Paper ID ICAPIE-2018-PE MT 218.

[22] Ali, Parvesh., Ranganath, M.S., Walia, R.S., Murtaza, Q.: Various developments in Abrasive flow machining process: A Review, 3rd international conference on Advanced Production and Industrial Engineering (ICAPIE'18), Paper ID ICAPIE-2018- PE MT 215.

[23] Bhardwaj, Anant., Ali, Parvesh., Walia, R. S., Murtaza, Qasim., Pandey, S. M.: Development of Hybrid Forms of Abrasive Flow Machining Process: A Review, Advances in Industrial and Production Engineering pp 41-67, Part of the Lecture Notes in Mechanical Engineering book series (LNME), conference paper (2019).

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/337739295

Multifunction Filter/Inverse Filter Configuration Employing CMOS CDBAs

Article · December 2019

CITATIONS
0

READS
58

3 authors:



Ram Bhagat
Delhi Technological University
7 PUBLICATIONS **7** CITATIONS
SEE PROFILE



Data Ram Bhaskar
Delhi Technological University
8 PUBLICATIONS **8** CITATIONS
SEE PROFILE



Pragati Kumar
Delhi Technological University
23 PUBLICATIONS **144** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Analog corcuit implementation of Fractional order filters View project

Novel OTA based inductorless realisation of Chaos's circuit View project

All content following this page was uploaded by Ram Bhagat on 05 December 2019.

The user has requested enhancement of the downloaded file.

Multifunction Filter/Inverse Filter Configuration Employing CMOS CDBAs

Ram Bhagat, D. R. Bhaskar, Pragati Kumar

Abstract: ***A voltage-mode (VM) multifunction configuration for the realization of conventional active filters and inverse active filters (IAF) using two current differencing buffered amplifiers and six passive elements has been presented. The proposed structure can realize low pass filter/inverse low pass filter (LPF/ILPF), high pass filter /inverse high pass filter (HPF/IHPF), and band pass filter/ Inverse band pass filter (BPF/IBPF) from the same circuit topology by appropriate choice(s) of the branch impedance(s). PSPICE simulations with CMOS current differencing buffered amplifiers implemented in 0.18µm CMOS TSMC technology have been presented to establish the workability of the proposed circuit configuration.***

Keywords: Analog Electronic Filters, Inverse Filters, Analog Signal Processing, Current Differencing Buffered Amplifier

I. INTRODUCTION

Current differencing buffered amplifier (CDBA) has drawn attention of researchers working in the field of analog signal processing and signal generation circuits as it provides ideally zero input impedance at its input terminals thus, eliminating the effect of parasitic elements at the input terminals. In addition, its low output impedance (ideally zero at the voltage output terminal) makes possible the cascading of CDBA-based circuit configurations without any loading effects. Numerous applications of CDBAs [1]-[17] in analog signal processing circuits have been presented by the researchers in open literature. Since this paper deals with the applications of CDBAs in realization of analog filters/inverse analog filters, in the following we present a brief overview of the important works done on the realization of filters/inverse filters using CDBAs to put the proposed work in this paper in proper prospective. In [1], single CDBA-based CM notch filter has been reported with two capacitors and two resistors. Realization of CM universal filter using three CDBAs, two capacitors and two resistors is reported in [2]. In [3], a multi-mode multifunction filter structure has been presented employing one CDBA, along with two/three resistors and capacitors. A third order VM all pass filter (APF) has been proposed using four CDBAs, three capacitors and eleven resistors in [4]. CM multipurpose filter configuration has been presented in [5] using single/two CDBA(s), three/four resistors and two/three capacitors. Two CDBA-based VM universal biquad filter and quadrature oscillator structure employing two capacitors and four resistors have been presented in [6]. Two electronically controllable configurations providing VM filters and quadrature sinusoidal oscillators using two CDBAs, one multiplier element, three capacitors and five resistors have been presented in [7]. A CM multifunction filter employing two CDBAs, three/four resistors, three/four capacitors is described in [8]. A single CDBA based multifunction biquad with three/four resistors and three/four capacitors has been reported in [9]. In [10] VM single CDBA-based universal filter has been presented with four resistors and two capacitors. A multiple output biquadratic filter has been reported in [11] using three CDBAs, two/four resistors and two/five capacitors. A CM KHN-biquad has been presented in [12] employing three CDBAs, two capacitors and six resistors. In [13], a VM single CDBA based APF and notch filter with three resistors and two capacitors has been reported. Two VM universal filter configurations with two CDBAs, four resistors and two capacitors have been described in [14]. It may be noted that in [13], and [14] the intrinsic property (current differencing) of the CDBA has not been used. Inverse analog active filters are an important class of analog signal processing circuits from the view point of some applications in the areas where the distortion introduced in the signal transmission path may be corrected by making the signal pass through an inverse filter whose characteristics are the reciprocal of the original system which has introduced the distortion. Over the years, several active inverse analog filter designs employing different active building blocks (ABBs) have been reported. Table I given below summarizes the salient features of active inverse analog filters.

Revised Manuscript Received on November 15, 2019
***** Correspondence Author
Ram Bhagat, Electrical Engineering Department, Delhi Technological University, Delhi, India.
D. R. Bhaskar*, Electronics and Communication Engineering Department, Delhi Technological University, Delhi, India.
Pragati Kumar, Electrical Engineering Department, Delhi Technological University, Delhi, India.

Table I. Comparison with earlier reported inverse filters

Ref.	Number of active devices used	Passive components used		Type of inverse filters realized	Whether conventional filters realized from the same circuit topology
		Resistor	Capacitor		
[15]	2 CDBA	2-4	2-4	ILPF,IHPF,IBPF,IBRF,IAPF	No
[16]	2 CDBA	3	3	ILPF,IHPF,IBPF	No
[17]	2 CDBA	4/5	2	IBRF, IAPF	No
[18]	1 NULLOR	4	2	IHPF	No
[19]	1 OA	3/7	2 (FCAP)	ILPF,IHPF,IBPF	No
[20]	3 CFOA	4	2	ILPF,IHPF,IBPF,IBRF	No
[21]	3 CFOA	3-5	2	ILPF,IHPF,IBPF,IBRF	No
[22]	3 CFOA	3	3	ILPF,IHPF,IBPF	No
[23]	3 CFOA	2-3	3-4	ILPF,IHPF,IBPF	No
[24]	2 CFOA	4-6	2	ILPF,IHPF,IBPF,IBRF	No
[25]	1 CCII	2	1	IAPF first order	No
[26]	4-6 CCII	3-4	2	ILPF, IHPF, IBPF	No
[27]	1 FTFN	5	2	ILPF	No
[28]	1FTFN	4	2	IAPF	No
[29]	1 FTFN	2-4	2/3	ILPF,IHPF,IBPF,IBRF,IAPF	No
[30]	2 OTRA	4	2/3	ILPF,IHPF,IBPF	No
[31]	2 OTRA	4-5	3-4	IBRF, IAPF	No
[32]	3 CDTA	2	2	IHPF	No
[33]	1 CDTA	1	1	IAPF first order	No
[34]	10/12 OTAs	0	2	ILPF,IHPF,IBPF	No
[35]	3 DDCC	2	2	ILPF,IHPF,IBPF	No
[36]	2/3/4 VDTA	0	2	ILPF,IHPF,IBPF,IBRF	No
Proposed	2CDBA	3	2	ILPF,IHPF,IBPF	Yes

FCAP: Fractional Capacitor

From the above table, it is observed that there is no configuration(s) reported earlier in the open literature which can realize conventional/inverse filters from the same topology with suitable choice(s) of passive circuit elements.

Thus, the main aim of this paper is to present a multifunction circuit topology with two CDBAs, four resistors and two capacitors (virtually grounded) to realize second-order conventional LPF, HPF, BPF and ILPF, IHPF and IBPF utilizing the intrinsic (current differencing) property of the CDBAs.

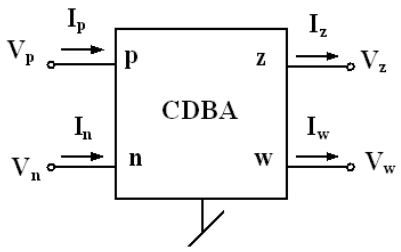


Fig. 1 Symbolic representation of CDBA

II. PROPOSED CIRCUIT CONFIGURATION

The CDBA symbol is shown in Fig.1 and its terminal voltage/current relationships are given in equation (1)

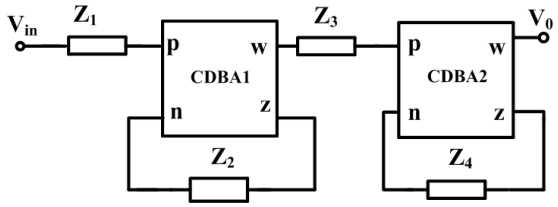


Fig. 2 The proposed circuit configuration

$$V_p = 0 = V_n,\ I_z = (I_p - I_n), \text{ and } V_w = V_z \quad (1)$$

The circuit for the realization of conventional LPF, HPF, BPF and ILPF, IHPF and IBPF with four branch impedances is shown in Fig. 2.

From a straight forward analysis, the transfer function (TF) for the configuration shown in Fig. 2 can be obtained as:

$$\frac{V_0}{V_{in}} = \frac{Z_2 Z_4}{4Z_1 Z_3} \quad (2)$$

By choosing suitable branch impedance(s), the various conventional filters and inverse filter TFs can be obtained as follows.

(A) Realization of conventional filters:

Case I: If we select

$Z_1 = R_1, Z_2 = \left(R_2 \| \frac{1}{sC_2}\right), Z_3 = R_3$ and $Z_4 = \left(R_4 \| \frac{1}{sC_4}\right)$, resulting TF becomes

$$\frac{V_0(s)}{V_{in}(s)} = \frac{\left(\frac{1}{4R_1R_3C_2C_4}\right)}{s^2 + \left(\frac{1}{R_4C_4} + \frac{1}{R_2C_2}\right)s + \frac{1}{R_2R_4C_2C_4}} \quad (3)$$

which represents a LPF. The cut-off frequency (ω_o), gain (H_o) and quality factor (Q) of this filter are given by:

$$\omega_o = \frac{1}{\sqrt{R_2R_4C_2C_4}},\ H_o = \frac{R_2R_4}{4R_1R_3} \text{ and } Q = \frac{\sqrt{R_2R_4C_2C_4}}{R_2C_2 + R_4C_4} \quad (4)$$

Case II: If the impedances are selected such that

$Z_1 = \left(R_1 + \frac{1}{sC_1}\right), Z_2 = R_2, Z_3 = \left(R_3 + \frac{1}{sC_3}\right)$ and $Z_4 = R_4$, *the* resulting TF becomes

$$\frac{V_0(s)}{V_{in}(s)} = \frac{\left(\frac{R_2R_4}{4R_1R_3}\right)s^2}{s^2 + \left(\frac{1}{4R_1C_1} + \frac{1}{R_3C_3}\right)s + \frac{1}{R_1R_3C_1C_3}} \quad (5)$$

which yields an HPF with ω_o, H_o and Q given by:

$$\omega_o = \frac{1}{\sqrt{R_1R_3C_1C_3}},\ H_o = \frac{R_2R_4}{4R_1R_3} \text{ and } Q = \frac{\sqrt{R_1R_3C_1C_3}}{R_1C_1 + R_3C_3} \quad (6)$$

Case III: (a) If we take

$Z_1 = \left(R_1 + \frac{1}{sC_1}\right), Z_2 = R_2, Z_3 = R_3$ and $Z_4 = \left(R_4 \| \frac{1}{sC_4}\right)$, then we get the transfer function of BPF given by:

$$\frac{V_0(s)}{V_{in}(s)} = \frac{\left(\frac{R_2}{4R_1R_3C_4}\right)s}{s^2 + \left(\frac{1}{R_1C_1} + \frac{1}{R_4C_4}\right)s + \frac{1}{R_1R_4C_1C_4}} \quad (7)$$

with ω_o, bandwidth (BW) and H_o given by:

$$\omega_o = \frac{1}{\sqrt{R_1R_4C_1C_4}};\ BW = \frac{1}{R_1C_1} + \frac{1}{R_4C_4} \text{ and } H_o = \frac{R_2R_4C_1}{4R_3(R_1C_1 + R_4C_4)} \quad (8)$$

(b) If we choose

$Z_1 = R_1, Z_2 = R_2, Z_3 = \left(R_3 + \frac{1}{sC_3}\right)$ and $Z_4 = \left(R_4 \| \frac{1}{sC_4}\right)$,

then we get the transfer function of an BPF given by:

$$\frac{V_0(s)}{V_{in}(s)} = \frac{\left(\frac{R_2}{4R_3R_1C_4}\right)s}{s^2 + \left(\frac{1}{R_3C_3} + \frac{1}{R_4C_4}\right)s + \frac{1}{R_3R_4C_3C_4}} \quad (9)$$

which gives ω_o, BW and H_o as:

$$\omega_o = \frac{1}{\sqrt{R_3R_4C_3C_4}};\ BW = \frac{1}{R_3C_3} + \frac{1}{R_4C_4} \text{ and } H_o = \frac{R_2R_4C_3}{4R_1(R_3C_3 + R_4C_4)} \quad (10)$$

(c) If we take

$Z_1 = \left(R_1 + \frac{1}{sC_1}\right), Z_2 = \left(R_2 \| \frac{1}{sC_2}\right), Z_3 = R_3$ and $Z_4 = R_4$, then we get the transfer function of an BPF given by:

$$\frac{V_0(s)}{V_{in}(s)} = \frac{\left(\frac{R_4}{R_1R_3C_2}\right)s}{s^2 + \left(\frac{1}{R_1C_1} + \frac{1}{R_2C_2}\right)s + \frac{1}{R_1R_2C_1C_2}} \quad (11)$$

with ω_o, BW and H_o given by:

$$\omega_o = \frac{1}{\sqrt{R_1R_2C_1C_2}};\ BW = \frac{1}{R_1C_1} + \frac{1}{R_2C_2} \text{ and } H_o = \frac{R_2R_4C_1}{4R_3(R_1C_1 + R_2C_2)} \quad (12)$$

(d) If we select $Z_1 = R_1, Z_2 = \left(R_2 \| \frac{1}{sC_2}\right), Z_3 = \left(R_3 + \frac{1}{sC_3}\right)$ and $Z_4 = R_4$, then we get the transfer function of an BPF given by:

$$\frac{V_0(s)}{V_{in}(s)} = \frac{\left(\frac{R_4}{R_3R_1C_2}\right)s}{s^2 + \left(\frac{1}{R_2C_2} + \frac{1}{R_3C_3}\right)s + \frac{1}{R_2R_3C_2C_3}} \quad (13)$$

which gives ω_o, BW and H_o as:

$$\omega_o = \frac{1}{\sqrt{R_2R_3C_2C_3}};\ BW = \frac{1}{R_2C_2} + \frac{1}{R_3C_3} \text{ and } H_o = \frac{R_2R_4C_3}{4R_1(R_2C_2 + R_3C_3)} \quad (14)$$

(B) Realization of ILPF, IHPF and IBPF responses:

Case I: If we select

$Z_1=\left(R_1\|\frac{1}{sC_1}\right),\ Z_2=R_2,\quad Z_3=\left(R_3\|\frac{1}{sC_3}\right)$ and $Z_4=R_4$ then

transfer function becomes

$$\frac{V_0(s)}{V_{in}(s)}=\frac{\left(\frac{4}{R_2R_4C_1C_3}\right)}{s^2+\left(\frac{1}{R_3C_3}+\frac{1}{R_1C_1}\right)s+\frac{1}{R_1R_3C_1C_3}} \quad (15)$$

which represents an ILPF with ω_o ,H_o and Q given by:

$$\omega_o=\frac{1}{\sqrt{R_1R_3C_1C_3}}\ ,H_o=\frac{4R_1R_3}{R_2R_4}\text{ and}\quad Q=\frac{\sqrt{R_1R_3C_1C_3}}{R_1C_1+R_3C_3} \quad (16)$$

Case II: If the impedances are selected, such that

$$Z_1=R_1,Z_2=\left(R_2+\frac{1}{sC_2}\right),\ Z_3=R_3,\text{ and }Z_4=\left(R_4+\frac{1}{sC_4}\right)$$

the transfer function becomes:

$$\frac{V_0(s)}{V_{in}(s)}=\frac{\left(\frac{4R_1R_3}{R_2R_4}\right)s^2}{s^2+\left(\frac{1}{R_2C_2}+\frac{1}{R_4C_4}\right)s+\frac{1}{R_2R_4C_2C_4}} \quad (17)$$

which gives an IHPF with ω_o, H_o and Q given by:

$$\omega_o=\frac{1}{\sqrt{R_2R_4C_2C_4}}\ ,\ H_o=\frac{4R_1R_3}{R_2R_4}\text{ and }Q=\frac{\sqrt{R_2R_4C_2C_4}}{R_2C_2+R_4C_4} \quad (18)$$

Case III: (a) If the impedances are selected, such that

$$Z_1=R_1,Z_2=\left(R_2+\frac{1}{sC_2}\right),\ Z_3=\left(R_3\|\frac{1}{sC_3}\right),\text{and }Z_4=R_4$$

then the transfer function realizes IBPF as

$$\frac{V_0(s)}{V_{in}(s)}=\frac{\left(\frac{4R_1}{R_2R_4C_3}\right)s}{s^2+\left(\frac{1}{R_2C_2}+\frac{1}{R_3C_3}\right)s+\frac{1}{R_2R_3C_2C_3}} \quad (19)$$

with ω_o, BW and H_o given by:

$$\omega_o=\frac{1}{\sqrt{R_2R_3C_2C_3}}\ ,\ BW=\frac{1}{R_2C_2}+\frac{1}{R_3C_3}\text{ and}\quad H_o=\frac{4R_1R_3C_2}{R_4(R_2C_2+R_3C_3)} \quad (20)$$

(b) If we take $Z_1=R_1,Z_2=R_2,\ Z_3=\left(R_3+\frac{1}{sC_3}\right)$ and $Z_4=\left(R_4\|\frac{1}{sC_4}\right)$,

then the transfer function represents the realization of IBPF

$$\frac{V_0(s)}{V_{in}(s)}=\frac{\left(\frac{4R_1}{R_2R_4C_3}\right)s}{s^2+\left(\frac{1}{R_3C_3}+\frac{1}{R_4C_4}\right)s+\frac{1}{R_3R_4C_3C_4}} \quad (21)$$

which realizes an IBPF with ω_o, BW and H_o given by:

$$\omega_o=\frac{1}{\sqrt{R_3R_4C_3C_4}}\ ,\ BW=\frac{1}{R_3C_3}+\frac{1}{R_4C_4}\text{ and}\quad H_o=\frac{4R_1R_3C_2}{R_4(R_2C_2+R_3C_3)} \quad (22)$$

(c) If we

take $Z_1=\left(R_1\|\frac{1}{sC_1}\right),Z_2=\left(R_2+\frac{1}{sC2}\right)Z_3=R_3,\text{and }Z_4=R_4,$ then the

resulting transfer function yields the realization of IBPF as

$$\frac{V_0(s)}{V_{in}(s)}=\frac{\left(\frac{4R_3}{R_2R_4C_1}\right)s}{s^2+\left(\frac{1}{R_2C_2}+\frac{1}{R_1C_1}\right)s+\frac{1}{R_2R_1C_2C_1}} \quad (23)$$

which realizes an IBPF with ω_o, BW and H_o given by:

$$\omega_o=\frac{1}{\sqrt{R_1R_2C_1C_2}}\ ,\ BW=\frac{1}{R_2C_2}+\frac{1}{R_1C_1}\text{ and}\quad H_o=\frac{4R_1R_3C_2}{R_4(R_2C_2+R_1C_1)} \quad (24)$$

(d) If we select

$Z_1=\left(R_1\|\frac{1}{sC_1}\right),Z_2=R_2,Z_3=R_3$ and $Z_4=\left(R_4+\frac{1}{sC_4}\right)$, then the

resulting transfer function represents the realization of IBPF

$$\frac{V_0(s)}{V_{in}(s)}=\frac{\left(\frac{4R_3}{R_2R_4C_1}\right)s}{s^2+\left(\frac{1}{R_1C_1}+\frac{1}{R_4C_4}\right)s+\frac{1}{R_4R_1C_4C_1}} \quad (25)$$

with ω_o, BW and H_o given by:

$$\omega_o=\frac{1}{\sqrt{R_1R_4C_1C_4}}\ ,\ BW=\frac{1}{R_1C_1}+\frac{1}{R_4C_4}\text{ and}\quad H_o=\frac{4R_1R_3C_4}{R_2(R_4C_4+R_1C_1)} \quad (26)$$

III. NON-IDEAL ANALYSIS

The real CDBA, in which the effects of different non-idealities associated with various terminals have been incorporated in terms of parasitic resistances and parasitic capacitances, is shown in Fig. 3.

Terminal equations of a real CDBA have been described [18] by the following relationships:

$$V_P=V_N=0,I_Z=\beta_PI_P-\beta_nI_n\text{and}V_W=\alpha V_Z \quad (27)$$

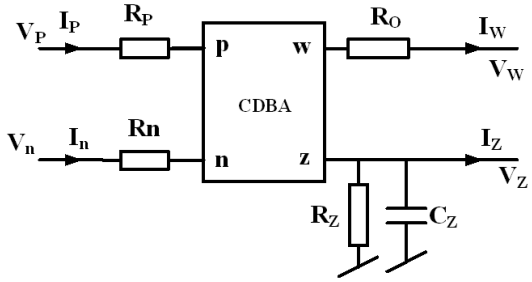


Fig. 3 Real CDBA model incorporating parasitic impedances

where $\beta_n = (1-\varepsilon_n)$ and $\varepsilon_n(|\varepsilon_n| << 1)$ is the current tracking error from n-terminal to z-terminal, $\beta_P = (1-\varepsilon_P)$ and $\varepsilon_p(|\varepsilon_p| << 1)$ is the current tracking error from p-terminal to z-terminal, and $\alpha = (1-\varepsilon_v)$ and $\varepsilon_v(|\varepsilon_v| << 1)$ is the voltage tracking error from *z*-terminal to w-terminal of the CDBA. In addition to the above, we have also considered parasitic resistance R_z, parasitic capacitance C_z at z-terminal of the CDBA in the non-ideal analysis given below and we get the following non-ideal transfer functions of conventional filters and inverse active filters:

(A) The various non-ideal expressions of LPF, HPF and BPF are given by:

$$\left.\frac{V_0(s)}{V_{in}(s)}\right|_{LPF} = \frac{\frac{1}{\alpha_1\alpha_2\beta_{p_1}\beta_{p_2}R_1R_3}}{As^2+Bs+C} \tag{28}$$

where $A = C_{z_1}C_{z_2} + (1+\beta_{n_1})(1+\beta_{n_2})C_2C_4$

$$B = \left(\begin{array}{l}\frac{C_{z2}}{R_{z2}} + \frac{C_{z1}}{R_{z1}} + \frac{C_{z1}(1+\beta_{n2})}{R_4} + \frac{C_{z2}(1+\beta_{n1})}{R_2} + \\ \left(\frac{C_2}{R_4} + \frac{C_4}{R_2}\right)(1+\beta_{n1})(1+\beta_{n2})\end{array}\right) \text{ and}$$

$$C = \left(\frac{1}{R_{z_1}R_{z_2}} + \frac{(1+\beta_{n2})}{R_{z_1}R_4} + \frac{(1+\beta n_1)}{R_{z_2}R_2} + \frac{(1+\beta_{n_1})(1+\beta_{n_2})}{R_2R_4}\right)$$

$$\left.\frac{V_0(s)}{V_{in}(s)}\right|_{HPF} = \frac{\frac{1}{\alpha_1\alpha_2\beta_{p_1}\beta_{p_2}R_1R_3}s^2}{As^4+Bs^3+Cs^2+Ds+E} \tag{29}$$

where

$A = C_{z_1}C_{z_2}$

$$B = \left(\frac{C_{z1}}{R_{z2}} + \frac{C_{z2}}{R_{z1}} + \frac{C_{z1}(1+\beta_{n2})}{R_4} + \frac{C_{z2}(1+\beta_{n1})}{R_2} + C_{z1}C_{z2}\left(\frac{1}{R_1C_1} + \frac{1}{R_3C_3}\right)\right)$$

$$C = \left(\begin{array}{l}\frac{1}{R_{z1}R_{z2}} + \frac{C_{z2}(1+\beta_{n1})}{R_2} + \frac{C_{z1}(1+\beta_{n2})}{R_4} + \frac{(1+\beta_{n2})}{R_4R_{z1}} + \frac{(1+\beta_{n1})}{R_2R_{z2}} + \\ \frac{(1+\beta_{n1})(1+\beta_{n2})}{R_2R_4} + \frac{C_{z1}}{R_{z2}}\left(\frac{1}{R_1C_1} + \frac{1}{R_3C_3}\right) + \\ \frac{C_{z2}}{R_{z1}}\left(\frac{1}{R_1C_1} + \frac{1}{R_3C_3}\right) + \frac{C_{z1}C_{z2}}{R_1R_3C_1C_3}\end{array}\right)$$

$$D = \left(\frac{1}{R_1C_1} + \frac{1}{R_3C_3}\right)\left(\frac{1}{R_{z1}R_{z2}} + \frac{(1+\beta_{n2})}{R_4R_{z1}} + \frac{(1+\beta_{n1})}{R_2R_{z2}} + \frac{(1+\beta_{n1})(1+\beta_{n2})}{R_2R_4}\right) + \frac{1}{R_1R_3C_1C_3}\left(\frac{C_{z1}}{R_{z2}} + \frac{C_{z2}}{R_{z1}} + \frac{C_{z2}(1+\beta_{n1})}{R_2} + \frac{C_{z1}(1+\beta_{n1})}{R_4}\right)$$

$$E = \frac{1}{R_1R_3C_1C_3}\left(\frac{1}{R_{z1}R_{z2}} + \frac{(1+\beta_{n2})}{R_4R_{z1}} + \frac{(1+\beta_{n1})}{R_2R_{z2}} + \frac{(1+\beta_{n1})(1+\beta_{n2})}{R_2R_4}\right)$$

$$\left.\frac{V_0(s)}{V_{in}(s)}\right|_{BPF} = \frac{\frac{1}{\alpha_1\alpha_2\beta_{P1}\beta_PR_1R_3}s}{As^3+Bs^2+Cs+D} \tag{30}$$

where

$A = C_{z1}C_{z2} + C_2C_{z2}(1+\beta_{n1})$

$$B = \left(\begin{array}{l}\frac{C_{z1}}{R_{z2}} + \frac{C_{z2}}{R_{z1}} + \frac{C_{z1}(1+\beta_{n2})}{R_4} + \frac{C_{z2}(1+\beta_{n1})}{R_2} + \frac{C_2(1+\beta_{n1})}{R_{z2}} + \frac{(1+\beta_{n1})(1+\beta_{n2})C_2}{R_4} + \\ \frac{C_{z1}C_{z2}}{R_1C_1} + \frac{C_2C_{z2}(1+\beta_{n1})}{R_1C_1}\end{array}\right)$$

$$C = \left(\begin{array}{l}\frac{1}{R_{z1}R_{z2}} + \frac{(1+\beta_{n2})}{R_4R_{z1}} + \frac{(1+\beta_{n1})}{R_2R_{z2}} + \frac{(1+\beta_{n1})(1+\beta_{n2})}{R_2R_4} + \frac{C_{z1}}{R_1C_1R_{z2}} + \frac{C_{z2}}{R_1C_1R_{z1}} + \\ \frac{C_{z2}(1+\beta_{n1})}{C_1R_1R_2} + \frac{C_{z1}(1+\beta_{n2})}{C_1R_1R_4} + \frac{C_2(1+\beta_{n1})}{C_1R_1R_{z2}} + \frac{(1+\beta_{n1})(1+\beta_{n2})C_2}{R_2R_4C_1}\end{array}\right)$$

$$D = \left(\frac{1}{R_1R_{z1}R_{z2}C_1} + \frac{(1+\beta_{n2})}{R_1R_4R_{z1}C_1} + \frac{(1+\beta_{n1})}{R_1R_2R_{z2}C_1} + \frac{(1+\beta_{n1})(1+\beta_{n2})}{R_1R_2R_4C_1}\right)$$

Since the parasitic resistance associated with the z terminal is very high (ideally infinite) and the parasitic capacitance associated with the z terminal is very small (ideally zero), we may approximate, the various filter transfer functions as:

$$\left.\frac{V_0(s)}{V_{in}(s)}\right|_{LPF} \cong \frac{\frac{\alpha_1\alpha_2\beta_{p1}\beta_{P2}}{C_2C_4R_3R_1(1+\beta_{n1})(1+\beta_{n2})}}{s^2+\left(\frac{1}{R_2C_2}+\frac{1}{R_4C_4}\right)s+\frac{1}{C_2C_4R_2R_4}} \tag{31}$$

which represents a LPF with non-ideal cut-off frequency ($\overline{\omega_0}$), gain ($\overline{H_0}$) and quality factor ($\overline{Q}$) given by

$$\begin{aligned}\overline{\omega_o} &= \frac{1}{\sqrt{R_2R_4C_2C_4}} = \omega_0, \overline{Q} = \frac{\sqrt{R_2R_4C_2C_4}}{R_2C_2+R_4C_4} = Q \\ \overline{H_o} &= \frac{\alpha_1\alpha_2\beta_{p1}\beta_{P2}R_2R_4}{(1+\beta_{n1})(1+\beta_{n2})R_1R_3} = \frac{4\alpha_1\alpha_2\beta_{p1}\beta_{P2}}{(1+\beta_{n1})(1+\beta_{n2})}H_0\end{aligned} \tag{32}$$

$$\left.\frac{V_0(s)}{V_{in}(s)}\right|_{HPF} \cong \frac{\frac{\alpha_1\alpha_2\beta_{p1}\beta_{P2}R_2R_4}{R_3R_1(1+\beta_{n1})(1+\beta_{n2})}s^2}{s^2+\left(\frac{1}{R_1C_1}+\frac{1}{R_3C_3}\right)s+\frac{1}{C_1C_3R_1R_3}} \tag{33}$$

which represents a HPF with $\overline{\omega_0}$, $\overline{H_0}$ and $\overline{Q}$ given by

$$\begin{aligned}\overline{\omega_o} &= \frac{1}{\sqrt{R_1R_3C_1C_3}} = \omega_0, \\ \overline{H_o} &= \frac{\alpha_1\alpha_2\beta_{p1}\beta_{P2}R_2R_4}{R_3R_1(1+\beta_{n1})(1+\beta_{n2})} = \frac{4\alpha_1\alpha_2\beta_{p1}\beta_{P2}}{(1+\beta_{n1})(1+\beta_{n2})}H_0 \\ \text{and } \overline{Q} &= \frac{\sqrt{R_1R_3C_1C_3}}{R_1C_1+R_3C_3} = Q\end{aligned} \tag{34}$$

$$\left.\frac{V_0(s)}{V_{in}(s)}\right|_{BPF} \cong \frac{\frac{1}{\alpha_1\alpha_2\beta_{P1}\beta_PR_1R_3}s}{s^2+\left(\frac{1}{R_1C_1}+\frac{1}{R_2C_2}\right)s+\frac{1}{C_1C_2R_1R_2}} \tag{35}$$

which represents a BPF with $\overline{\omega_0}$, $\overline{H_0}$ and $\overline{BW}$ given by

$$\begin{aligned}\overline{\omega_o} &= \frac{1}{\sqrt{R_1R_2C_1C_2}} = \omega_0 \text{ ; } \overline{BW} = \frac{1}{R_2C_2} + \frac{1}{R_1C_1} = BW \\ \text{and } \overline{H_o} &= \frac{4\alpha_1\alpha_2\beta_{p1}\beta_{p2}}{(1+\beta_{n1})(1+\beta_{n2})}H_0\end{aligned} \tag{36}$$

From equations (32), (34) and (36) it is noted that the non-ideal values of cut-off frequency, quality factor and bandwidth of the conventional

LPF, HPF and BPF do not depend on voltage/current tracking error coefficients and parasitic elements associated with various terminals of the CDBA, however the gain of the filter depends on the tracking error parameters.

(B) Similarly, the various non-ideal transfer functions of ILPF, IHPF and IBPF are given by:

$$\left.\frac{V_0(s)}{V_{in}(s)}\right|_{ILPF} = \frac{1}{\dfrac{C_{z1}C_{z2}}{\alpha_1\alpha_2\beta_{p1}\beta_{p2}C_1C_3}s^2 + Bs + C}{s^2 + \dfrac{1}{\alpha_1\alpha_2\beta_{p1}\beta_{p2}}\left(\dfrac{1}{R_1C_1}+\dfrac{1}{R_3C_3}\right)s + \dfrac{1}{\alpha_1\alpha_2\beta_{p1}\beta_{p2}C_1C_3R_1R_3}} \quad (37)$$

where

$$B = \left(\frac{C_{z1}}{R_{z2}} + \frac{C_{z2}}{R_{z1}} + \frac{C_{z1}}{R_4}(1+\beta_{n2}) + \frac{C_{z2}}{R_2}(1+\beta_{n1})\right)\frac{1}{\alpha_1\alpha_2\beta_{P1}\beta_{p2}C_1C_3}$$

$$C = \frac{1}{\alpha_1\alpha_2\beta_{P1}\beta_{P2}}\left(\frac{1}{R_{z1}R_{z2}C_1C_3} + \frac{(1+\beta_{n2})}{C_1C_3R_4R_{z1}} + \frac{(1+\beta_{n1})}{C_1C_3R_2R_{z2}} + \frac{(1+\beta_{n1})(1+\beta_{n2})}{C_1C_3R_4R_2}\right)$$

$$\left.\frac{V_0(s)}{V_{in}(s)}\right|_{IHPF} = \frac{1}{As^4 + Bs^3 + Cs^2 + Ds + E}{s^2 + \dfrac{1}{\alpha_1\alpha_2\beta_{p1}\beta_{P2}}\left(\dfrac{1}{R_2C_2}+\dfrac{1}{R_4C_4}\right)s + \dfrac{1}{\alpha_1\alpha_2\beta_{p1}\beta_{P2}R_2R_4C_2C_4}} \quad (38)$$

where

$$A = \frac{C_{z1}C_{z2}R_1R_3}{\alpha_1\alpha_2\beta_{P1}\beta_{P2}}$$

$$B = \frac{1}{\alpha_1\alpha_2\beta_{P1}\beta_{P2}}\left(\begin{array}{l} R_1R_3C_{z1}C_{z2} + \dfrac{C_{z1}C_{z2}R_1R_3}{C_4R_4} + \dfrac{C_{z2}R_1R_3}{R_{Z1}} + \dfrac{C_{Z1}R_1R_3}{R_4}(1+\beta_{n2}) + \\ \dfrac{C_{Z2}R_1R_3}{R_2}(1+\beta_{n2}) \end{array}\right)$$

$$C = \frac{R_1R_3}{\alpha_1\alpha_2\beta_{P1}\beta_{P2}}\left(\begin{array}{l} \dfrac{1}{R_{z1}R_{Z2}} + \dfrac{1}{C_1R_2R_Z} + \dfrac{1}{C_2R_2}\left(\dfrac{C_{z1}}{R_{z2}}+\dfrac{C_{z2}}{R_{z1}}\right) + \dfrac{1}{C_4R_4}\left(\dfrac{C_{z1}}{R_{z2}}+\dfrac{C_{z2}}{R_{z1}}\right) + \\ \dfrac{C_{z1}C_{z2}}{C_2C_4R_2R_4} + \dfrac{C_{z1}}{C_2R_2R_4}(1+\beta_{n2}) + \dfrac{C_{z2}}{C_2R_2R_4}(1+\beta_{n1}) + \\ \dfrac{1}{R_4R_{z1}}(1+\beta_{n2}) + \dfrac{1}{R_2R_{z2}}(1+\beta_{n1}) + \dfrac{1}{R_4R_2}(1+\beta_{n2})(1+\beta_{n1}) \end{array}\right)$$

$$D = \frac{R_1R_3}{\alpha_1\alpha_2\beta_{P1}\beta_{P2}}\left(\begin{array}{l} \dfrac{1}{R_{z1}R_{Z2}R_2} + \dfrac{1}{C_4R_4R_{Z1}R_{z2}} + \dfrac{1}{C_2C_4R_2R_4}\left(\dfrac{C_{z1}}{R_{z2}}+\dfrac{C_{z2}}{R_{z1}}\right) + \dfrac{1+\beta_{n2}}{C_2R_2C_4R_4} \\ + \dfrac{1+\beta_{n1}}{C_2R_2C_4R_4} \end{array}\right)$$

and $E = \dfrac{R_1R_3}{\alpha_1\alpha_2\beta_{P1}\beta_{P2}C_2C_4R_2R_4R_{z1}R_{z2}}$

$$\left.\frac{V_0(s)}{V_{in}(s)}\right|_{IBPF} = \frac{1}{As^3 + Bs^2 + Cs + D}{s^2 + \dfrac{1}{\alpha_1\alpha_2\beta_{p1}\beta_{P2}}\left(\dfrac{1}{R_2C_2}+\dfrac{1}{R_3C_3}\right)s + \dfrac{1}{\alpha_1\alpha_2\beta_{p1}\beta_{P2}R_2R_3C_2C_3}} \quad (39)$$

where

$$A = \frac{C_{z1}C_{z2}R_1}{\alpha_1\alpha_2\beta_{P1}\beta_PC_3}$$

$$B = \frac{R_1}{\alpha_1\alpha_2\beta_{P1}\beta_{P2}C_3}\left(\frac{C_{z1}}{R_{z2}} + \frac{C_{z2}}{R_{z1}} + \frac{C_{z1}}{R_4}(1+\beta_2) + \frac{C_{z2}}{R_2}(1+\beta_{n1}) + \frac{C_{z2}C_{z1}}{R_2C_2}\right)$$

$$C = \frac{R_1}{\alpha_1\alpha_2\beta_{P1}\beta_{P2}C_3}\left(\begin{array}{l} \dfrac{1}{R_{z1}R_{Z2}} + \dfrac{1}{C_2R_2}\left(\dfrac{C_{z1}}{R_{z2}}+\dfrac{C_{z2}}{R_{z1}}\right) + \dfrac{C_{z1}}{C_2R_2R_4}(1+\beta_{n2}) + \\ \dfrac{(1+\beta_{n1})}{R_2R_{z2}} + \dfrac{(1+\beta_{n2})(1+\beta_{n1})}{R_4R_2} \end{array}\right)$$

$$D = \frac{R_1}{\alpha_1\alpha_2\beta_{P1}\beta_{P2}C_3}\left(\frac{1}{R_{z1}R_{z2}R_2C_2} + \frac{1+\beta_{n2}}{C_2R_2R_4R_{z1}}\right)$$

Again, as R_z is very high (ideally infinite) and the value of C_z is very low (ideally zero), we may approximate, the various inverse filter transfer functions as:

$$\left.\frac{V_0(s)}{V_{in}(s)}\right|_{ILPF} \cong \frac{1}{\dfrac{(1+\beta_{n1})(1+\beta_{n2})}{\alpha_1\alpha_2\beta_{P1}\beta_{P2}C_1C_3R_4R_2}}{s^2 + \dfrac{1}{\alpha_1\alpha_2\beta_{P1}\beta_{P2}}\left(\dfrac{1}{R_1C_1}+\dfrac{1}{R_3C_3}\right)s + \dfrac{1}{\alpha_1\alpha_2\beta_{P1}\beta_{P2}C_1C_3R_1R_3}} \quad (40)$$

which represents an ILPF with $\overline{\omega_0}$, $\overline{H_0}$ and $\overline{Q}$ given by:

$$\overline{\omega_0} = \frac{1}{\sqrt{\alpha_1\alpha_2\beta_{P1}\beta_{P2}C_1C_3R_1R_3}} = \frac{\omega_0}{\sqrt{\alpha_1\alpha_2\beta_{P1}\beta_{P2}}}$$

$$\overline{H_0} = \frac{R_1R_3(1+\beta_{n1})(1+\beta_{n2})}{R_2R_4} = \frac{(1+\beta_{n1})(1+\beta_{n2})H}{4} \quad (41)$$

$$\overline{Q} = \frac{\sqrt{\alpha_1\alpha_2\beta_{P1}\beta_{P2}C_1C_3R_1R_3}}{R_1C_1 + R_3C_3} = \sqrt{\alpha_1\alpha_2\beta_{P1}\beta_{P2}}Q$$

$$\left.\frac{V_0(s)}{V_{in}(s)}\right|_{IHPF} \cong \frac{1}{\dfrac{R_1R_3(1+\beta_{n1})(1+\beta_{n2})}{\alpha_1\alpha_2\beta_{P1}\beta_{P2}R_2R_4}s^2}{s^2 + \dfrac{1}{\alpha_1\alpha_2\beta_{p1}\beta_{P2}}\left(\dfrac{1}{R_2C_2}+\dfrac{1}{R_4C_4}\right)s + \dfrac{1}{\alpha_1\alpha_2\beta_{p1}\beta_{P2}C_2C_4R_2R_4}} \quad (42)$$

which realizes IHPF with $\overline{\omega_0}$, $\overline{H_0}$ and $\overline{Q}$ given by:

$$\overline{\omega_0} = \frac{1}{\sqrt{\alpha_1\alpha_2\beta_{P1}\beta_{P2}C_2C_4R_2R_4}} = \frac{\omega_0}{\sqrt{\alpha_1\alpha_2\beta_{P1}\beta_{P2}}}$$

$$\overline{H_0} = \frac{R_1R_3(1+\beta_{n1})(1+\beta_{n2})}{\alpha_1\alpha_2\beta_{P1}\beta_{p2}R_2R_4} = \frac{(1+\beta_{n1})(1+\beta_{n2})H}{4} \quad (43)$$

$$\overline{Q} = \frac{\sqrt{\alpha_1\alpha_2\beta_{P1}\beta_{P2}C_2C_4R_2R_4}}{R_2C_2 + R_4C_4} = \sqrt{\alpha_1\alpha_2\beta_{P1}\beta_{P2}}Q$$

$$\left.\frac{V_0(s)}{V_{in}(s)}\right|_{IBPF} \cong \frac{1}{\dfrac{R_1R_3(1+\beta_{n1})(1+\beta_{n2})}{\alpha_1\alpha_2\beta_{P1}\beta_{P2}R_2R_4}s}{s^2 + \dfrac{1}{\alpha_1\alpha_2\beta_{p1}\beta_{P2}}\left(\dfrac{1}{R_2C_2}+\dfrac{1}{R_3C_3}\right)s + \dfrac{1}{\alpha_1\alpha_2\beta_{p1}\beta_{P2}R_2R_3C_2C_3}} \quad (44)$$

which realizes IBPF with $\overline{\omega_0}$ and $\overline{BW}$ given by:

$$\overline{\omega_0} = \frac{1}{\sqrt{\alpha_1\alpha_2\beta_{p1}\beta_{P2}R_2R_3C_2C_3}} = \frac{\omega_0}{\sqrt{\alpha_1\alpha_2\beta_{p1}\beta_{P2}}},$$

$$\overline{BW} = \frac{1}{\alpha_1\alpha_2\beta_{p1}\beta_{P2}}\left(\frac{1}{R_2C_2}+\frac{1}{R_3C_3}\right) = \frac{1}{\alpha_1\alpha_2\beta_{p1}\beta_{P2}}BW \quad (45)$$

Equations (41), (43) and (45) reveal that the cutoff frequency, gain, quality factor and bandwidth of the ILPF, IHPF and IBPF are affected by voltage/current tracking errors of the CDBA terminals.

IV. SENSITIVITY ANALYSIS

Various sensitivities (both active as well as passive) of ω_o, Q and *BW* of conventional filters (LPF, HPF, BPF) and inverse active filters (ILPF, IHPF, IBPF) have been determined and are presented in Table II. It may be observed that the sensitivities with respect to active and passive components are less than $\pm$ 1.

Table II. Sensitivity of conventional filters/ inverse filters

Filter	Sensitivity
LPF	$S_{R2}^{\omega_O} = S_{R4}^{\omega_O} = -\frac{1}{2}, S_{C2}^{\omega_0} = S_{C4}^{\omega_O} = -\frac{1}{2}$, $S_{R_1}^{\omega_0} = S_{R_3}^{\omega_0} = 0$, $S_{\alpha 1}^{\omega_O} = S_{\alpha 2}^{\omega 0} = S_{\beta p1}^{\omega_O} = S_{\beta p2}^{\omega_O} = 0 = S_{\beta n_1}^{\omega_0} = S_{\beta n2}^{\omega_0} = 0$ $S_{R2}^{Q} = S_{C2}^{Q} = \frac{R_4C_4}{2(C_2R_2 + R_4C_4)}, S_{R4}^{Q} = S_{C4}^{Q} = \frac{R_2C_2}{2(C_2R_2 + R_4C_4)}$, $S_{\alpha 1}^{\omega_O} = S_{\alpha 2}^{\omega 0} = S_{\beta p1}^{\omega_O} = S_{\beta p2}^{\omega_O} = 0 = S_{\beta n_1}^{\omega_0} = S_{\beta n2}^{\omega_0} = 0$
HPF	$S_{R_1}^{\omega_O} = S_{R3}^{\omega_O} = -\frac{1}{2}, S_{C1}^{\omega_0} = S_{C3}^{\omega_O} = -\frac{1}{2}$, $S_{R_2}^{\omega_0} = S_{R_4}^{\omega_0} = 0$ $S_{\alpha 1}^{\omega_O} = S_{\alpha 2}^{\omega 0} = S_{\beta p1}^{\omega_O} = S_{\beta p2}^{\omega_O} = 0$, $S_{\beta n_1}^{\omega_0} = S_{\beta n2}^{\omega_0} = 0$ $S_{R1}^{Q} = S_{C1}^{Q} = \frac{R_3C_3}{2(C_1R_1 + R_3C_3)}, S_{R3}^{Q} = S_{C3}^{Q} = \frac{R_1C_1}{2(C_1R_1 + R_3C_3)}$, $S_{\alpha 1}^{\omega_O} = S_{\alpha 2}^{\omega 0} = S_{\beta p1}^{\omega_O} = S_{\beta p2}^{\omega_O} = 0 = S_{\beta n_1}^{\omega_0} = S_{\beta n2}^{\omega_0} = 0$
BPF Case(a)	$S_{R_1}^{\omega_O} = S_{R4}^{\omega_O} = -\frac{1}{2}, S_{C1}^{\omega_0} = S_{C4}^{\omega_O} = -\frac{1}{2}$, $S_{R_2}^{\omega_0} = S_{R_4}^{\omega_0} = 0$, $S_{\alpha 1}^{\omega_O} = S_{\alpha 2}^{\omega 0} = S_{\beta p1}^{\omega_O} = S_{\beta p2}^{\omega_O} = 0$, $S_{\beta n_1}^{\omega_0} = S_{\beta n2}^{\omega_0} = 0$ $S_{R1}^{BW} = S_{C1}^{BW} = -\frac{R_4C_4}{C_1R_1 + R_4C_4}, S_{R4}^{BW} = S_{C4}^{BW} = -\frac{R_1C_1}{C_1R_1 + R_4C_4}$, $S_{\alpha 1}^{\overline{BW}} = S_{\alpha 2}^{\overline{BW}} = S_{\beta p1}^{\overline{BW}} = S_{\beta p2}^{\overline{BW}} = 0$
ILPF	$S_{R_1}^{\omega_O} = S_{R3}^{\omega_O} = -\frac{1}{2}, S_{C1}^{\omega_0} = S_{C3}^{\omega_O} = -\frac{1}{2}$, $S_{R_2}^{\omega_0} = S_{R_4}^{\omega_0} = 0$, $S_{\alpha 1}^{\omega_O} = S_{\alpha 2}^{\omega 0} = S_{\beta p1}^{\omega_O} = S_{\beta p2}^{\omega_O} = -\frac{1}{2}$, $S_{\beta n_1}^{\omega_0} = S_{\beta n2}^{\omega_0} = 0$ $S_{R1}^{Q} = S_{C1}^{Q} = \frac{R_3C_3}{2(C_1R_1 + R_3C_3)}, S_{R3}^{Q} = S_{C3}^{Q} = \frac{R_1C_1}{2(C_1R_1 + R_3C_3)}$, $S_{\alpha 1}^{\omega_O} = S_{\alpha 2}^{\omega 0} = S_{\beta p1}^{\omega_O} = S_{\beta p2}^{\omega_O} = \frac{1}{2} = S_{\beta n_1}^{\omega_0} = S_{\beta n2}^{\omega_0} = 0$
IHPF	$S_{R_2}^{\omega_O} = S_{R_4}^{\omega_O} = S_{C4}^{\omega_O} = S_{C_2}^{\omega_O} = -\frac{1}{2}$, $S_{R_1}^{\omega_0} = S_{R_3}^{\omega_0} = 0$ $S_{\alpha 1}^{\omega_O} = S_{\alpha 2}^{\omega_O} = S_{\beta p1}^{\omega_O} = S_{\beta p2}^{\omega_O} = -\frac{1}{2}$, $S_{\beta n_1}^{\omega_0} = S_{\beta n2}^{\omega_0} = 0$, $S_{R2}^{Q} = S_{C2}^{Q} = \frac{R_4C_4}{2(C_2R_2 + R_4C_4)}, S_{R4}^{Q} = S_{C4}^{Q} = \frac{R_2C_2}{2(C_2R_2 + R_4C_4)}$, $S_{\alpha 1}^{\omega_O} = S_{\alpha 2}^{\omega 0} = S_{\beta p1}^{\omega_O} = S_{\beta p2}^{\omega_O} = \frac{1}{2} = S_{\beta n_1}^{\omega_0} = S_{\beta n2}^{\omega_0} = 0$
IBPF Case(a)	$S_{R_2}^{\omega_O} = S_{R_3}^{\omega_O} = S_{C3}^{\omega_O} = S_{C_2}^{\omega_O} = -\frac{1}{2}$, $S_{R_1}^{\omega_0} = S_{R_4}^{\omega_0} = 0$, $S_{\alpha 1}^{\omega_O} = S_{\alpha 2}^{\omega_O} = S_{\beta p1}^{\omega_O} = S_{\beta p2}^{\omega_O} = -\frac{1}{2}$, $S_{\beta n_1}^{\omega_0} = S_{\beta n2}^{\omega_0} = 0$ $S_{R2}^{BW} = S_{C2}^{BW} = -\frac{R_3C_3}{C_3R_3 + R_2C_2}, S_{R3}^{BW} = S_{C3}^{BW} = -\frac{R_1C_1}{C_1R_1 + R_4C_4}$, $S_{\alpha 1}^{\overline{BW}} = S_{\alpha 2}^{\overline{BW}} = S_{\beta p1}^{\overline{BW}} = S_{\beta p2}^{\overline{BW}} = -1$

V. SIMULATION RESULTS

The workability of the multifunction filter configuration has been verified using PSPICE simulations wherein the employed CDBA structure shown in Fig. 4 has been derived from the structure of a current differencing current conveyor [38]. The bias voltages of ±2.5V and the bias current of 40µA were used in the simulations. The aspect ratios used for implementation of CDBA are given in Table III. The parameters selected for a cutoff frequency of 1.59 MHz are as follows.

LPF: $C_1 = C_3 = 10pF$, $R_1 = R_3 = 5k\Omega$, $R_2 = R_4 = 10K\Omega$

HPF: $C_1 = C_3 = 10pF$, $R_1 = R_3 = 10k\Omega$, $R_2 = R_4 = 20k\Omega$

BPF (case a): $C_3 = C_4 = 10pF$, $R_1 = R_2 = R_4 = 10k\Omega$, $R_3 = 5k\Omega$

ILPF: $C_1 = C_3 = 10pF$, $R_1 = R_3 = 10k\Omega$, $R_2 = R_4 = 20K\Omega$

IHPF: $C_2 = C_4 = 10pF$, $R_1 = R_3 = 5k\Omega$, $R_2 = R_4 = 10k\Omega$

IBPF (case a): $C_2 = C_3 = 10pF$, $R_1 = R_2 = R_3 = 10k\Omega$, $R_4 = 20k\Omega$

Fig. 5 shows the frequency responses of conventional LPF, HPF and BPF and Fig. 6 shows the frequency responses of ILPF, IHPF and IBPF. The input signal level was kept at 0.5V peak to peak. These results establish the workability of the proposed configuration.

Table III. aspect ratio of MOSFETs used for realization of CDBA

MOSFETs	W:L
M_1,M_2,M_{17},M_{23},M_{25}	50 um:0.5um
M_3,M_4,M_{18},M_{24},M_{26}	100um:0.5um
M_5-M_{16},M_{19}-M_{22}	3.33um:0.5um

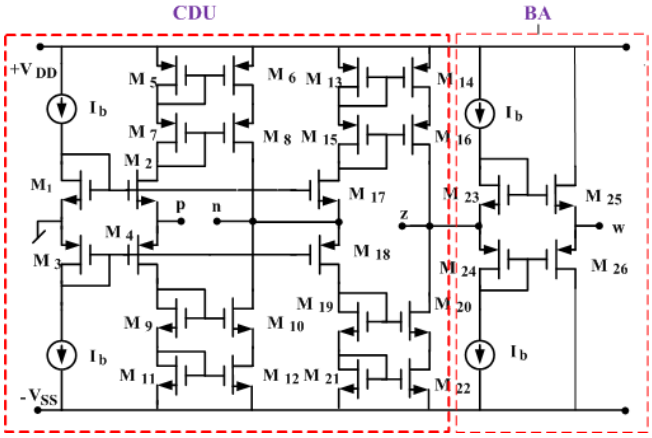
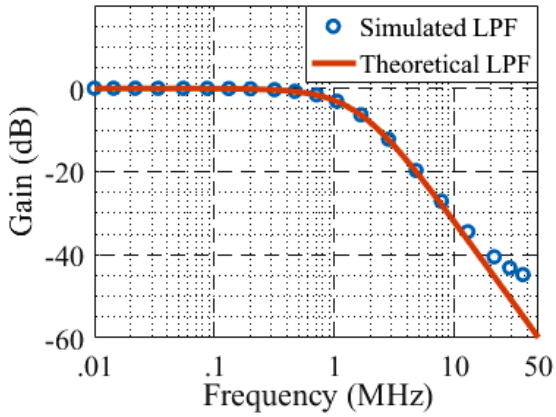
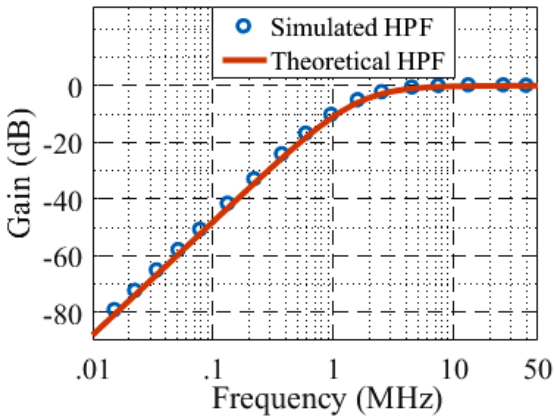


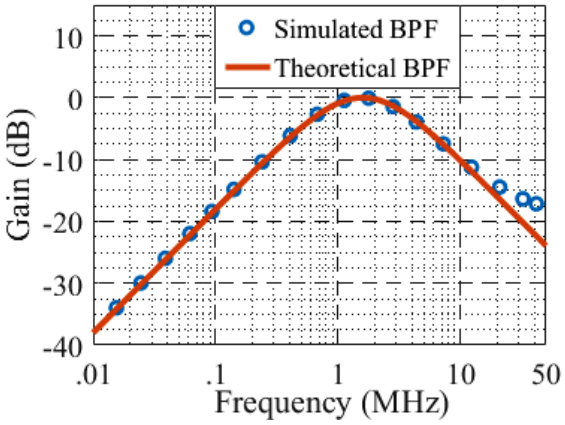
Fig. 4. CMOS realization of CDBA



(a)

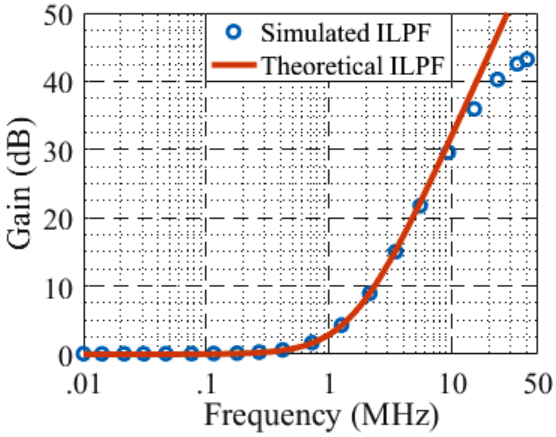


(b)

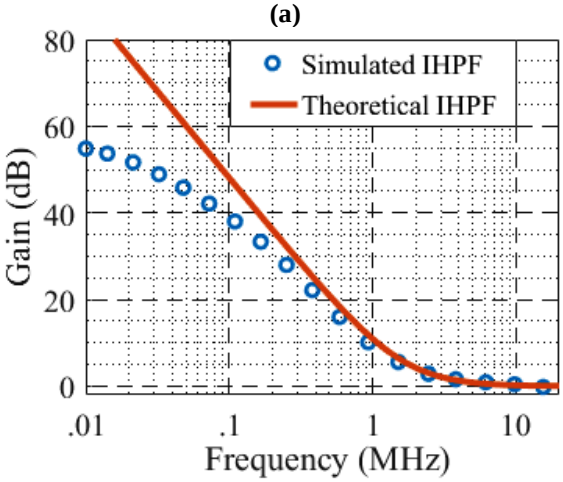


(c)

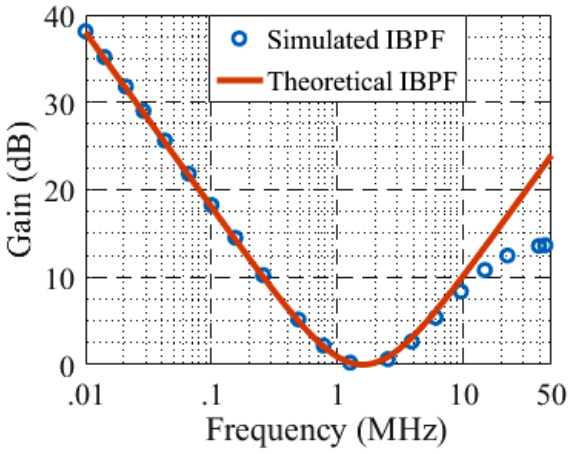
Fig. 5 Simulated frequency responses for conventional filters (a) LPF, (b) HPF and (c) BPF



(a)



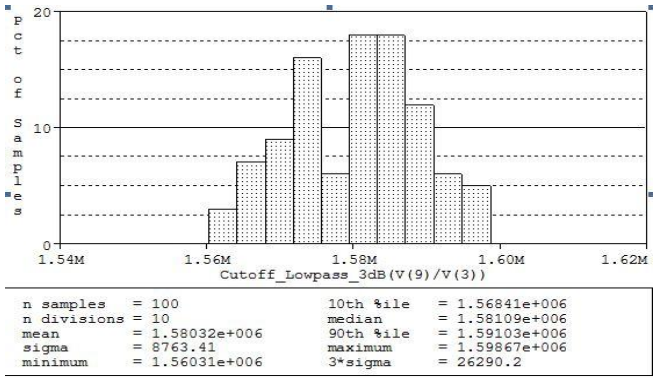
(b)



(c)

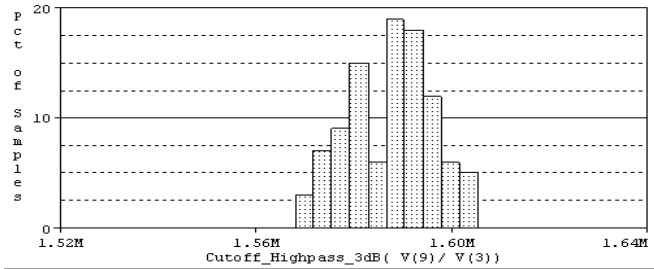
Fig. 6 Simulated frequency responses of inverse filters (a) ILPF (b) IHPF and (c) IBPF

The Monte-Carlo simulations of the conventional filters and inverse active filters showing variation of cut off frequency and gain with 5 percent variation in capacitance and resistance values have been carried out and the results have been shown in Fig.7(a-f).



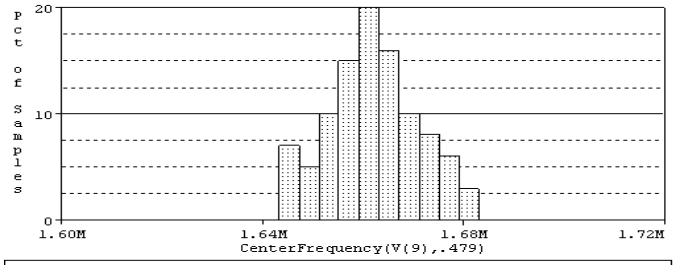
n samples	= 100	10th %ile	= 1.56841e+006
n divisions	= 10	median	= 1.58109e+006
mean	= 1.58032e+006	90th %ile	= 1.59103e+006
sigma	= 8763.41	maximum	= 1.59867e+006
minimum	= 1.56031e+006	3*sigma	= 26290.2

(a) Variation in cut-off frequency of LPF with 5 percent variation in resistance and capacitance value



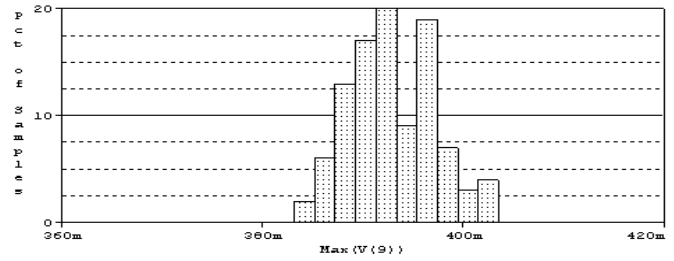
n samples	= 100	10th %ile	= 1.57587e+006
n divisions	= 10	median	= 1.58799e+006
mean	= 1.58744e+006	90th %ile	= 1.59794e+006
sigma	= 8586.67	maximum	= 1.60534e+006
minimum	= 1.56795e+006	3*sigma	= 25760

(b) Variation in cut-off frequency of HPF with 5% variation in resistance and capacitance value



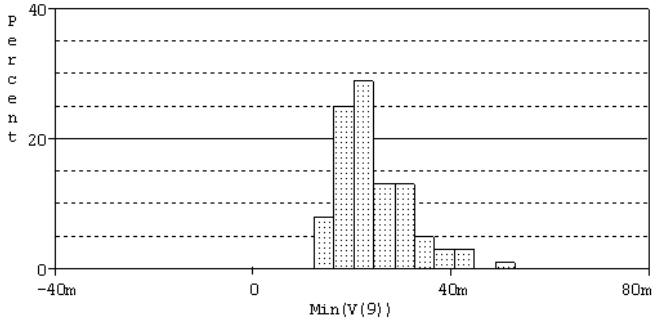
n samples	= 100	10th %ile	= 1.64976e+006
n divisions	= 10	median	= 1.66158e+006
mean	= 1.66197e+006	90th %ile	= 1.67474e+006
sigma	= 9022.99	maximum	= 1.68318e+006
minimum	= 1.64326e+006	3*sigma	= 27069

(c) Variation in centre frequency of BPF with 5% variation in resistance and capacitance value



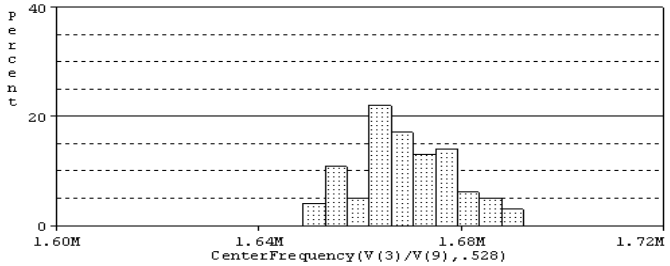
n samples	= 100	10th %ile	= 0.387503
n divisions	= 10	median	= 0.392739
mean	= 0.392957	90th %ile	= 0.398306
sigma	= 0.00431561	maximum	= 0.403561
minimum	= 0.383144	3*sigma	= 0.0129468

(d) Variation in gain of ILPF with 5% variation in resistance and capacitance value



n samples	= 100	10th %ile	= 0.016663
n divisions	= 10	median	= 0.0229539
mean	= 0.0242919	90th %ile	= 0.0341688
sigma	= 0.00721257	maximum	= 0.0529833
minimum	= 0.0122693	3*sigma	= 0.0216377

(e) Variation in cut-off frequency of IHPF with 5% variation in resistance and capacitance value



n samples	= 100	10th %ile	= 1.65576e+006
n divisions	= 10	median	= 1.66902e+006
mean	= 1.66896e+006	90th %ile	= 1.68232e+006
sigma	= 9680.5	maximum	= 1.69236e+006
minimum	= 1.64879e+006	3*sigma	= 29041.5

(f) Variation in centre frequency of IBPF with 5% variation in resistance and capacitance value

Fig. 7 Monte-Carlo simulation results of the (a) LPF (b) HPF (c) BPF (d) ILPF (e) IHPF and (f) IBPF

VI. CONCLUSION

A multifunction structure employing two CDBAs and six passive components has been presented. The proposed structure can realize conventional as well as inverse active filter responses by appropriate selection(s) of branch impedance(s) as resistor(s)/capacitor(s). The active and passive sensitivities of the presented filters are low.Monte-Carlo analysis has also been carried out to check the robustness of the described circuit. The workability of the proposed filter has been validated by PSPICE and MATLAB simulation employing an exemplary CMOS CDBA [38] architecture.

REFERENCES

1. C. Cakir, O. Cicekoglu, “Low-voltage high-performance CMOS current differencing buffered amplifier (CDBA)”, Proceedings of IEEE PRIME Conference, Istanbul, 22 June 2008-25 April 2008, pp. 37-40. http://dx.doi.org/10.1109/rme.2008.4595719.
2. V. Sawangarom, T. Dumawipata, W. Tangsrirat, W. Surakampontorn, “Cascadable three-input single-output current-mode universal filter using CDBAs”, In The 2007 ECTI International Conference, 2007, pp. 53-56.
3. M. Koksal, S. E. Oner, M. Sagbas, “A new second-order multi-mode multi-function filter using a single CDBA”, In 2009 European Conference on Circuit Theory and Design , 2009, pp. 699-702
4. C. Acar, S. Ozoguz, (1999), “A new versatile building block: current differencing buffered amplifier suitable for analog

signal-processingfilters", Microelectronics Journal, 1999, 30(2), pp. 157-160.
5. S. Özcan, H. Kuntman, O. Çiçekolu, "Cascadable current mode multipurpose filters employing current differencing buffered amplifier (CDBA)", AEU-International Journal of Electronics and Communications, 2002, 56(2), pp. 67-72.
6. W. Tangsrirat, T. Pukkalanun, W. Surakampontorn, "CDBA-based universal biquad filter and quadrature oscillator", Active and Passive Electronic Components, 2008.
7. R. Nandi, P. Venkateswaran, S. Das, M. Kar, "CDBA-based electronically tunable filters and sinusoid quadrature oscillator", Journal of telecommunication, 2010, 4(1), pp. 35-41.
8. A. Ü. Keskin, E. Hancioglu, "Current mode multifunction filter using two CDBAs", AEU-International Journal of Electronics and Communications, 2005, 59(8), pp. 495-498.
9. A. Ü. Keskin, "Multi-function biquad using single CDBA", Electrical Engineering, 2002, 88, pp. 353-356.
10. S. A. Bashir, N. A. Shah, "Voltage mode universal filter using current differencing buffered amplifier as an active device.", Circuits and Systems, 2012, 3(03), pp. 278.
11. W. Tangsrirat, W. Surakampontorn, "Realization of multiple-output biquadratic filters using current differencing buffered amplifiers", International Journal of Electronics, 2005, 92(6), pp.313-325.
12. A. Toker, S. Özoğuz, C. Acar, "Current-mode KHN-equivalent biquad using CDBAs", Electronics Letters, 1999, 35(20), pp.1682-1683.
13. C. Cakir, S. Minaei, O. Cicekoglu, "Low voltage low power CMOS current differencing buffered amplifier", Analog Integrated Circuits and Signal Processing, 2010, 62(2), pp. 237-244.
14. J. K. Pathak, A. K. Singh, R. Senani, "New voltage mode universal filters using only two CDBAs", ISRN Electronics, 2013.
15. R. Pandey, N. Pandey , T. Negi, V. Garg , "CDBA based universal inverse filter", ISRN Electronics, 2013.
16. A.R. Nasir, S. N. Ahmad, "A new current-mode multifunction inverse filter using CDBAs", International Journal of Computer Science and Information Security, 2013, 11(12), pp. 50.
17. R. Bhagat, D. R. Bhaskar, P. Kumar, "Inverse Band Reject and All Pass Filter Structure Employing CMOS CDBAs", Int. J. of Engineering Research and Technology, 2019, 08(09).
18. Leuciuc, "Using nullors for realization of inverse transfer functionsand characteristics", Electro. Lett., 33(11) (1997), 949–951.
19. D. R. Bhaskar, M. Kumar, and P. Kumar, "Fractional order inverse filters using operational amplifier", Analog Integr. Circuits and Signal Process., 97(2018), 149-158.
20. S. S. Gupta, D. R. Bhaskar, and R. Senani, A.K. Singh, " Inverse active filters employing CFOA", Elect. Eng. 1(2009), 23-26.
21. S. S. Gupta, D.R. Bhaskar, and R. Senani, "New analogue inverse filters realized with current feedback op-amp", Int. j. of Electro.9(2011),1103–1113.
22. H. Y. Wang, S.H. Chang, T.Y. Yang, and P.Y. Tsai, "A novel multifunction CFOA based inverse filter", Circuits and Syst.2(2011),14–17.
23. K. Garg, R. Bhagat, and B. Jaint, "A novel multifunction modified CFOA based inverse filter", In Power Electronics (IICPE), IEEE 5th India International Conference. (2012), 1-5.
24. V. N. Patil, and R. K. Sharma, "Novel inverse active filters employing CFOA", Int. J. for Scientific Research & Develop. 3(2015), 359–360.
25. N. A. Shah, and M. F. Rather, "Realization of voltage-mode CCII-based all pass filter and its inversion version", Indian J. Pure & Applied Physics, 44(2006), 269–271.
26. T. Tsukutani, Y. Kunugasa, and N. Yabuki, "CCII-Based Inverse Active Filters with Grounded Passive Components", J. Electr. Eng. 6(2018):212-215.
27. B. Chipipop, and W. Surakampontorn, "Realization of current-mode FTFN-based inverse filter", Electron. Lett. 35(1999), 690–692.
28. H. Y. Wang, and C.T. Lee, "Using nullors for realisation of current-mode FTFN based inverse filters", Electron. Lett. 35(1999),1889–1890.
29. M. T. Abuelma'atti, "Identification of cascadable current-mode filters and inverse-filters using single FTFN", Frequenz , 54(11–12)(2000), 284–289.
30. A. K. Singh, A. Gupta, and R. Senani, "OTRA-based multi-function inverse filter configuration", Adv. in Elect. and Electron. Eng. 15(2018), 846-856.
31. A. Pradhan, and R. K. Sharma, "Generation of OTRA-Based Inverse All Pass and Inverse Band Reject Filters", Proceedings of the National Acad. of Sci. India Section A: Physical Sciences, 1-11 (2019)https://doi.org/10.1007/s40010-019-00603-w
32. A. Sharma, A. Kumar, and P. Whig, "On performance of CDTA based novel analog inverse low pass filter using 0.35 µm CMOS parameter", Int. J. of Sc., Tech. & Manag.4(2015), 594–601.
33. N. A. Shah, M. Quadri, and S. Z. Iqbal, "High output impedance current-mode all pass inverse filter using CDTA", Indian J. of pure and app. physics,46(2018), 893–896.
34. T. Tsukutani , Y. Sumi, and N. Yabuki, "Electronically tunable inverse active filters employing OTAs and grounded capacitors", Int. J. of Electron. Lett.4(2016), 166–176.
35. N. Herencsar , A. Lahiri, J. Koton, and K. Vrba, "Realizations of second-order inverse active filters using minimum passive components and DDCCs", In Proceedings of 33rd Int. Conference on Telecomm. and Signal Proc.-TSP, (2010), 38-41.
36. P. Kumar, N. Pandey, and S.K. Paul, "Realization of Resistor less and Electronically Tunable Inverse Filters Using VDTA", J. Circuits, Syst. Comp. (2018),1950143.
37. J. K. Pathak, A. K. Singh, R. Senani, "Systematic realisation of quadrature oscillators using current differencing buffered amplifiers", IET circuits, devices & systems, 5(3), 2011, pp. 203-211.
38. A. K. Singh, P. Kumar, " A novel fully differential current mode universal filter", In Circuits and Systems (MWSCAS),2014 IEEE 57th International Midwest Symposium, 2014: pp. 579-582.

AUTHORS PROFILE



Ram Bhagat Associate Professor in Department of Electrical Engineering, Delhi Technological University, New Delhi, India. Currently, he is pursuing Ph.D. from the Delhi Technological University. His research interests are in the area of Network Analysis and Synthesis, Microelectronics, CMOS Analog Integrated Circuits and Control System.



D. R. Bhaskar received B.Sc. degree from Agra University, B. Tech. degree from Indian Institute of Technology (IIT) Kanpur, M. Tech. from IIT Delhi and Ph.D. from University of Delhi. Prof. Bhaskar held the positions of Lecturer (1984–1990) and Senior Lecturer (1990–1995) at the Electrical Engineering Department of Delhi College of Engineering (now Delhi Technological University). He joined the Electronics and Communication Engineering (ECE) Department of Jamia Millia Islamia in July 1995, as a Reader and became a Professor in January 2002. He served as the Head of the Department of ECE from 2002 to 2005. Presently, he is working in the Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, India. His teaching and research interests are in the areas of Bipolar and CMOS Analog Integrated Circuits and Systems, Current Mode/Voltage Mode Signal Processing, Communication Systems, Fractional Order Filters and Electronic Instrumentation. Prof. Bhaskar has authored or co-authored 92 research papers-all in international journals of repute, 9 international conference papers and 4 book chapters. He has co-authored 3 monographs published by Springer. He has acted/has been acting as a Reviewer for several journals of IEEE, IEE and other international journals of repute.



Pragati Kumar Professor in Department of Electrical Engineering, Delhi Technological University, New Delhi, India was born in Muzaffarpur, Bihar, India. He received his Ph.D. from the University of Delhi in 2008. His teaching and research interests are in the area of Network Analysis and Synthesis, Microelectronics, CMOS Analog Integrated Circuits and Fractional Order Circuits/Systems. He has co-authored 13 research papers in various international journal/conferences/edited book volumes.

View publication stats

Proc. of the Eighth Intl. Conf. on Advances in Computing, Communication and Information Technology - CCIT 2019
Copyright © Institute of Research Engineers and Doctors. All rights reserved.
ISBN: 978-1-63248-169-6 DOI : 10.15224/978-1-63248-169-6-09

Multiple Resource Management and Burst Time Prediction using Deep Reinforcement Learning

Vaibhav Kumar, Siddhant Bhambri, Prashant Giridhar Shambharkar
Delhi Technological University
India

***Abstract*—Resource management and job scheduling are two problems that go hand-in-hand and the solutions to which are primarily dependent on the nature of workload. With increasing demand to automate the entire process from allocating resources to scheduling jobs efficiently, deep reinforcement learning techniques have been brought into the picture which adapt to the environment and learn from experience. In this paper, we present SchedQRM which classifies burst time of jobs based on their signature and employs Deep Q-Network algorithm to find an optimal solution for any arbitrary job set. We also evaluate our proposed work against state-of-the-art heuristics to show the efficacy of our approach.**

***Keywords*— reinforcement learning, job scheduling , Deep-Q Network**

Introduction

Resource management has always been a tricky domain in the field of research and has become increasingly significant due to the rapid developments in production technologies. The problem of optimal allocation and use of resources has been dealt in the past in several ways [1-5]. Improvement in the classical measures of efficiency due to periodic rescheduling has already been addressed in the past alongside the undesirable effect of compromising stability.

Considering the heuristics on which Reinforcement Learning (RL) algorithms work, we believe that RL approaches and methodologies fit well in the domain of resource management and job scheduling since they shall allow the machine to check for the best possible order of scheduling for a set of jobs, given the resource and burst time requirements for each job.

Our approach is an extension to the idea of machines being able to handle resources on their own in a justified manner. RL has gained attention in the field of machine learning research. The concept of decision-making had been introduced earlier in the research problems of resource management and job scheduling. RL revolves around agents which interact with the environment to accomplish a task. For each action it takes during its discourse through the environment, the agent receives a reward -positive or negative, based on the result of the action it takes. The agent has no prior knowledge of the task to be performed and learns based on the reward it receives.

We design and evaluate SchedQRM, an online multi-resource job scheduler, in our approach to applying RL for solving the problem of resource management. A set of jobs are fed into the scheduler as an input along with their job signature, and no pre-emption is allowed. The scheduler aims to optimize average job slowdown or job completion time by minimizing it.

Related Work

Resource allocation problem has been addressed in various contexts such as in Radio Networks [1], Software Defined Networking [2], mobile cloud computing networks [3] and in wireless communication networks [4]. Wan et al. [5] in their paper propose a resource allocation algorithm to maximize throughput for hybrid Visible Light Communication (VLC) and Wi-Fi networks. In all these papers, we observe that increase in the throughput, whatever the requirement may be, has been the primary objective.

The objectives while scheduling jobs/tasks at hand vary in context from minimizing CPU energy [6] to reducing total completion time on a single machine [7]. With the advent of Big Data and Machine Learning approaches, Karim H. and Ahmed J. in [8] proposed an approach for scheduling tasks in Big Data Cluster and showed a comparison with the traditional task schedulers such as the First-In-First-

SEEK DIGITAL LIBRARY

Proc. of the Eighth Intl. Conf. on Advances in Computing, Communication and Information Technology - CCIT 2019
Copyright © Institute of Research Engineers and Doctors. All rights reserved.
ISBN: 978-1-63248-169-6 DOI : 10.15224/978-1-63248-169-6-09

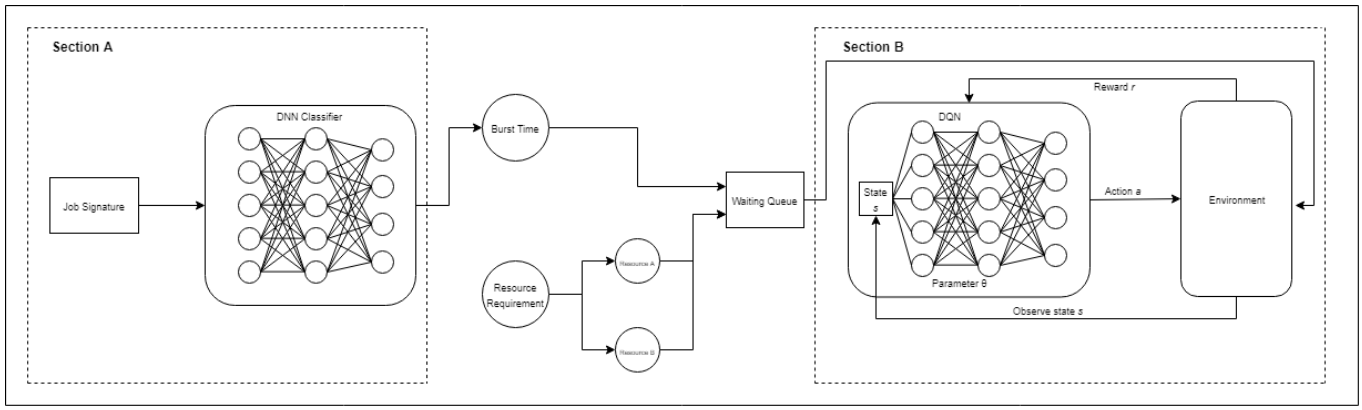


Figure 1: Flow of data in our proposed model.

Out (FIFO) scheduler and the Capacity Scheduler. Fuming et al. in [9] proposed the concept of a virtual scheduling pool whereas Zhao P. and Huang T. in [10] incorporated a genetic algorithm to solve the problem of single resource dynamic Job-Shop scheduling. Smart manufacturing domain also faces with a real-time requirement of job scheduling which has been tackled using a hybrid computing framework [11].

Resource Isolation Policy (RIP) combined with static as well as dynamic scheduling strategy was proposed by Liu et al. [12] to solve the problem of hard real-time task deadline. In [13], the authors characterize the performance of scheduling policies for wireless systems that are based on Cumulative Density Functions (CDF). Su N. et al. [14] incorporated genetic programming to propose an automatic design of scheduling policies which shows outstanding performance on unseen simulation scenarios. Particle Swarm Optimization (PSO) [15] algorithm has also been used for optimizing task scheduling in the field of cloud computing.

Survey on the past work is done which shows the use of reinforcement learning approach to designing feedback controllers for discrete as well as dynamic systems [16]. In [17], the authors proposed an adaptive Neural Net-based controller using RL for a class of nonlinear systems which does not require information about the system dynamics. RL has also been used to solve the problem of resource allocation [18] where the authors combine the strengths of RL and queuing models in a hybrid approach. However, our inspiration has been from the work of Hongzi M. et al. in [19] where the state space has been represented pictorially and fed into a deep reinforcement learning network to find the optimal scheduling policy for a given job-set.

Background

This section discusses the techniques in brief that we have worked upon in this paper.

Burst Time Classification: Prior knowledge of the burst time of a job helps exceptionally well in resource allocation and job scheduling. In a few cases, the burst time (run time) of a job is known, but mostly an approximation needs to be made. We divide the burst time of every job into a certain number of classes based on the job environment. Every job has a signature/set of attributes. These are fed to a neural net classifier to classify jobs and approximate the burst time.

Reinforcement Learning: Consider a scenario where there is an *agent* which interacts with an *environment*. The *agent* observes a state s_t and chooses an action a_t at each time step t, from the set of possible actions. Once the action is taken, a state transitions takes place from s_t to s_{t+1}, following which a reward r_t is given to the *agent*. The state transitions and rewards are assumed to have the Markov property; i.e., the action to be taken by the *agent* in the current state s_t is independent of the states that preceded s_t.

Note that the *agent* has no prior knowledge of which state of the *environment* would it transition to or what reward it may receive, once it chooses to take action a_t. It is while interacting with the

SEEK DIGITAL LIBRARY

Proc. of the Eighth Intl. Conf. on Advances in Computing, Communication and Information Technology - CCIT 2019
Copyright © Institute of Research Engineers and Doctors. All rights reserved.
ISBN: 978-1-63248-169-6 DOI : 10.15224/978-1-63248-169-6-09

environment, during training, that the *agent* will be able to observe the value of these quantities. The expected cumulative discounted reward: $E[\sum_{t=0}^{\infty} \gamma^t r_t]$, where $\gamma \in (0,1]$ is the discount factor for future rewards, needs to be maximized through learning.

Deep Q-Network (DQN): DQN is a type of Temporal Difference (TD) learning method. With the use of TD learning methods, the estimate of the final reward calculated at every step for each state can be formally expressed as:

$$\boldsymbol{V(s_t) \leftarrow V(s_t) + \alpha[r_{t+1} + \gamma V(s_{t+1}) - V(s_t)]}. \quad (1)$$

Where, $V(s_t)$ represents the utility value of state s_t, $\boldsymbol{\alpha}$ is the learning rate, $\boldsymbol{\gamma}$ is the discount factor and $\boldsymbol{r_{t+1}}$ is the observed reward at time $\boldsymbol{t+1}$. Compared to Monte Carlo methods [20], where the Q values are updated after the end of an episode; here, the Q values are updated after ever action. This helps in guiding the *agent* to the goal state more efficiently. The *agent* uses DQN because of the advantages that it offers in solving the scheduling problem through Experience Replay. Experience Replay is a circular queue which stores agent's experiences in form a tupple $e_t=(s_t, a_t, r_t, s_{t+1})$. Here, the *agent* takes an action a_t to move from state s_t to s_{t+1} and gains a reward r_t. This helps reducing correlation between transitions when the neural network has to be updated. The learning speed of the model increases with mini-batches that are made by DQN to update the neural network being employed. It also reuses past transitions to avoid catastrophic forgetting which speeds up learning and also breaks undesirable temporal correlations. Thus, DQN is believed to achieve stable training.

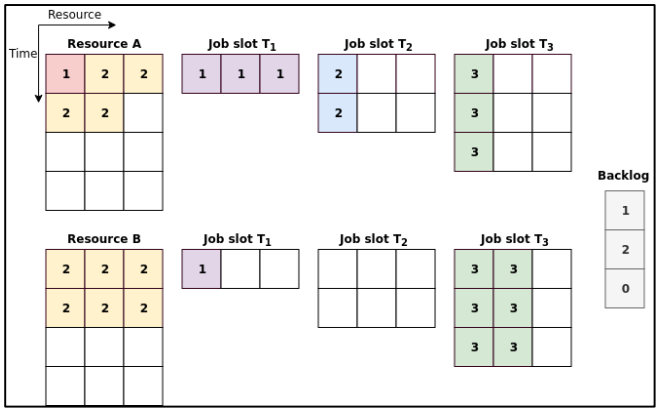


Figure 2: Pictorial representation of a sample state in SchedQRM.

Design

In this section, we present the design of SchedQRM. We describe the problem and also its formulation as an RL task. We then explain our solution to this problem based on the techniques described in the previous section.

Model

As shown in Figure 1, we divide our model into 2 sections- Section A and Section B.

Section A takes job signature as an input and predicts the burst time for the job using a Deep Neural Network (DNN). Resource requirements along with the burst time for the job are sent as an input to the waiting queue.

Section B consists of the DQN model which outputs a scheduling policy for a given job set. All the jobs from the waiting queue are fed into the *environment* as a starting state.

We consider a cluster with k resource types (e.g. CPU, I/O, memory) and it is treated as a single collection of resources. Jobs arrive at the cluster in discrete timestamps. One or more of the waiting jobs are chosen to be scheduled by the scheduler at each timestamp. The resource requirement of each job j is given by the vector $\boldsymbol{r_j = (r_1, r_2, \ldots\ldots, r_k)}$, where r_i represents the number of instances of resource i required by the job and $\boldsymbol{i \in [0, k]}$. We define T_j as the duration/execution time of the job. Given the above information, the DNN correctly places each job into a class from the set $\boldsymbol{(c_1, c_2, \ldots\ldots, c_n)}$, where c_i represents a class of jobs

SEEK DIGITAL LIBRARY

Proc. of the Eighth Intl. Conf. on Advances in Computing, Communication and Information Technology - CCIT 2019
Copyright © Institute of Research Engineers and Doctors. All rights reserved.
ISBN: 978-1-63248-169-6 DOI : 10.15224/978-1-63248-169-6-09

that requires resources for *i* timestamps. The jobs are assumed to be non-preemptive for the sake of simplicity. Also, r_j must be allocated to the job *j* continuously from the time that the job starts execution until completion.

The simplicity of the model can allow it to be used in other domains of application where a similar type of input is provided containing a set of jobs along with their resource requirements. Given this information, our design of the model will identify the execution time requirement of the job by learning from the experience. This information shall further be passed to the DQN *agent* which will schedule these jobs.

TABLE I: INPUT DATASET TO THE NEURAL NET CLASSIFIER

	interp	gnu.hash	dynsym	dynstr	rela.dyn	rela.pit	pit	text	rodata	eh_frame_hdr	eh_frame	got	got.pit	bss	symtab	strtab	filesize	time
0	28	0	5	2	0	4	1	645	32064	1	0	0	1	0	1	764	49768	1.0
1	28	0	5	2	0	4	1	645	32072	1	0	0	1	0	1	764	49768	1.0
2	28	0	5	2	0	4	1	645	32080	1	0	0	1	0	1	764	49768	1.0

Objective: Similar to the prior work shown in [19], we use the *average job slowdown* as the primary objective for our agent. Formally, the slowdown for each job *i* is given by $\boldsymbol{S_j = C_j/T_j}$, where C_j is the completion time and T_j is the burst time of the job. Completion time of a job is the time between arrival and completion of execution; note that $\boldsymbol{S_j \geq 1}$. If the completion time of the job is not normalized by the job's duration, the solution will be biased towards large jobs.

Burst Time Classificaion

Dataset creation: To create the dataset for burst time prediction, we use C++ object files for four programs to generate job signatures, namely: Matrix multiplication, Quicksort, Fibonacci series generator, and a random number generator. A script is deployed to create Executable and Linkable Format (ELF) files with various input sizes, and it collects the job signatures through the *readelf* bash command. We run this script until 100,000 data points are created. File size and run time are also stored for every ELF file along with the signature. Table I represents 3 out of the total data points/rows of the data-set that are fed into the neural network classifier for training. This dataset was created on an Intel core i7-6700 quad-core, 64-bit x86 processor and 8+8GB DDR4 3000mhz Corsair RAM.

RL formation

State Space: We represent a pictorial representation of a single state of the system as shown in Figure 2. This state contains information regarding the current resources' allocation, the jobs in the waiting queue and the jobs present in the *backlog*. The left-most clusters represent the allocation of resource instances to jobs which have been scheduled for service as of the current timestep. Here we have assumed two types of resources with three instances available for each of them. The resource allocation shown is present starting from the first timestamp till *t* timestamps, each row representing a timestamp. Jobs in the waiting queue belong to one of the time slots belonging to the vector $\boldsymbol{T_i = (T_1, T_2, \ldots\ldots, T_n)}$, where *i* refers to the number of timestamps required by a job to complete its execution; for example, the job in time slot T_2 requires one instance of resource A and zero instances of resource B for two timestamps. The different numbers within the resource clusters represent different jobs belonging to the respective time slots that have already been assigned resources and are undergoing or are about to begin execution; for example, 2 represents that a job belonging to T_2 has currently been assigned two instances of Resource A and three instances of Resource B for two timestamps. The first job of a certain burst time is represented in it's appropriate job slot while others wait for their turn in the *backlog*. The t^{th} box in the backlog stores the number of jobs with burst time *t*. For example, in Figure 2, there is one job of burst time 1 and two jobs of burst time 2 in the *backlog*.

Our state representation is a modified version of the state representation shown in [19], unlike which, we have a fixed representation of jobs based on their burst time. This fixed representation allows us

SEEK DIGITAL LIBRARY

Proc. of the Eighth Intl. Conf. on Advances in Computing, Communication and Information Technology - CCIT 2019
Copyright © Institute of Research Engineers and Doctors. All rights reserved.
ISBN: 978-1-63248-169-6 DOI : 10.15224/978-1-63248-169-6-09

to represent a state as an array of numbers and obviates the need for pictorial representation of a state and involvement of lofty convolutional neural networks. Hence, our input to the model is a flattened array representation of Figure 2.

Note: By having a fixed time representation, only a single job of a particular burst time can be represented in a state. Multiple jobs of same burst time have to wait in the waiting queue which might hinder the learning of the *agent*. However, it helps significantly in an optimized representation of any arbitrary job set which makes our algorithm very robust.

Action Space: We choose the action space to be simple, and it is given by $a_t \in \{1, 2,..,i,.., N\}$, where N is the maximum burst time and $a_t = i$ means that the *agent* should schedule the job at the i^{th} slot, which, because of our fixed time state representation, has a burst time of *i*. A valid decision is one in which the *agent* chooses to schedule a job at the non-empty slot. An invalid decision is the one where the *agent* selects an action corresponding to an empty slot as it makes no sense to schedule a job which does not exist. Once the *agent* takes a valid decision, a job is scheduled in the first possible timestamp of the resource clusters in which the resource requirements of the job can be completely satisfied till completion. A state transition is then observed: the scheduled job is allocated it's appropriate position in the resource clusters.

Rewards: Since DQN is a TD learning method, we have crafted a dynamic reward system that will guide our *agent* to the optimal policy by giving an appropriate reward at every time step. We do this by maintaining a counter *c* for the current time in the *environment*. If the *agent* decides to take action a_t, then it is given a reward $r_t = -(a_t + c/a_t)$. If no job exists at the chosen time slot, then a very high negative reward is given. We set the discount factor $\gamma = 1$ so that the cumulative reward of an episode equates the negative of the sum of job slowdown. This way, maximizing the cumulative reward over an episode is equivalent to minimizing the average slowdown.

Evaluation

We evaluate SchedQRM to answer the following questions:

1. How accurately does the DNN time classifier predict the burst time of incoming jobs?
2. How does SchedQRM compare with state-of-the-art heuristics when scheduling online jobs having multiple resource requirement?

Classifier Training and Testing: Before training the *agent* for job scheduling, we train a DNN classifier over the job signature to predict the burst time. Rather than using burst time as a continuous variable, we classify it into ten equally spaced classes. This classification helps significantly in our RL formulation. Data points with outliers and large burst times are discarded to keep the classes

TABLE II: CLASSIFICATION REPORT FOR DNN TIME CLASSIFIER

	Precision	**Recall**	**F1-Score**
Class 1	1.00	0.99	1.00
Class 2	0.95	0.99	0.97
Class 3	0.99	0.96	0.97
Class 4	0.94	0.93	0.93
Class 5	0.89	0.86	0.88
Class 6	0.88	0.95	0.91
Class 7	0.94	0.90	0.92
Class 8	0.92	0.98	0.95
Class 9	0.98	0.81	0.89
Class 10	0.83	1.00	0.90

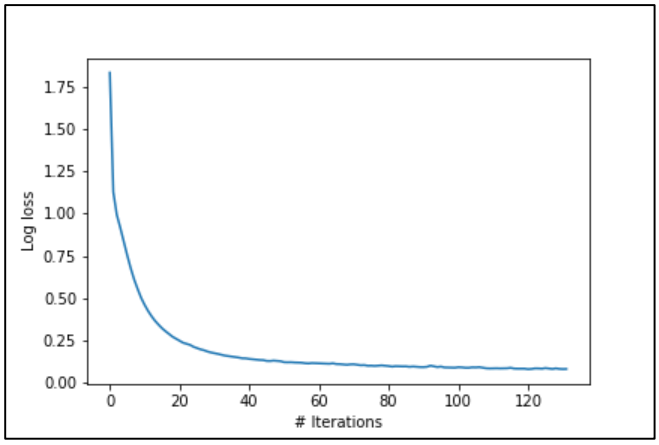


Figure 3: Training loss curve of the classifier.

SEEK DIGITAL LIBRARY

Proc. of the Eighth Intl. Conf. on Advances in Computing, Communication and Information Technology - CCIT 2019
Copyright © Institute of Research Engineers and Doctors. All rights reserved.
ISBN: 978-1-63248-169-6 DOI : 10.15224/978-1-63248-169-6-09

balanced. A simple DNN classifier with 100 hidden layers, 10 output layers and a learning rate of 0.0001 is trained over 70,000 data points. Rest of 30,000 data points are used to test our model. Our model efficiently converges after 120 iterations by using Adam optimizer [22].

Given the signature of a job, the classifier can predict the burst time class of that job with high accuracy, and this can be seen from Table II. The table shows the values of three evaluation metrics we have calculated for each class 1-10, namely- Precision, Recall, and F1-Score. The classifier can yield satisfying results for each of the classes. Figure 3 shows the training loss curve of the classifier and it's convergence.

DQN Training and Testing: We create a complete job scheduling environment with custom states, actions, and rewards. Our *agent* explores this *environment* and learns an optimal policy with the help of two neural networks and a *replay memory* as used in the DQN algorithm. To implement Experience Replay, we have used a circular memory buffer called *replay memory* which stores new transitions by overwriting the previous ones. The purpose of using a replay memory is:

1. Remembering experience: By storing and sampling transitions from experience replay, our *agent* gets exposed to a broader set of experience and knowledge that helps the *agent* learn more efficiently.
2. De-correlation: If we merely train our *agent* in sequential order, we risk getting our *agent* influenced by the correlation between consecutive states. By randomly sampling transitions from the experience replay, we enable learning from an independent and identically distributed experience.

Job signatures are picked at random from the dataset and fed to the system in an online fashion. This way SchedQRM is trained for arbitrary jobsets and is expected to optimally schedule any set of jobs. which makes it very robust. The cluster load or the number of jobs selected are varied as a percentage of the number of time classes from 10% (1 job for 10-time classes) to 200% (20 jobs for 10-time classes). This set of job is then passed to the DNN classifier to estimate the time of each job. On random, for every job, a dominant or both equally dominant resources are chosen.

In the case of dominant resource, it's resource requirement is chosen between 50% and 100% of maximum resource instances while for the other case, it is chosen in between 0% and 50%. In the case of equally dominant resources, the resource requirement for both the resources is chosen randomly between 10%-100% of total instances. Such job sets are then used to train the SchedQRM.

Every new job set is loaded into the environment's initial state. DQN trains on these job sets to finally converge and form an optimal policy which can be used to determine a scheduling policy for any new arbitrary job set. As stated earlier, we have used two neural networks namely the *evaluate* network and the *train* network for our learning algorithm.

The weights of the *train* network are transferred to the *evaluate* network after every 1000 timestamps, and this helps in stabilizing the DQN algorithm. Each neural network has a single fully connected hidden layer and 20 output layers, one each for an action. The replay memory is a buffer of length 2000. We set the learning rate $\alpha = 0.01$, $\varepsilon = 0.9$ and $\boldsymbol{\gamma} = 1$ for training our *agent*.

Figure 4 represents the plot of the burst time of a job belonging to one of the classes from Class 1 to Class 10 versus the average job slowdown time measured over different jobsets. The extended line shows the maximum slowdown time that was observed for a job belonging to any class.

SEEK DIGITAL LIBRARY

Proc. of the Eighth Intl. Conf. on Advances in Computing, Communication and Information Technology - CCIT 2019
Copyright © Institute of Research Engineers and Doctors. All rights reserved.
ISBN: 978-1-63248-169-6 DOI : 10.15224/978-1-63248-169-6-09

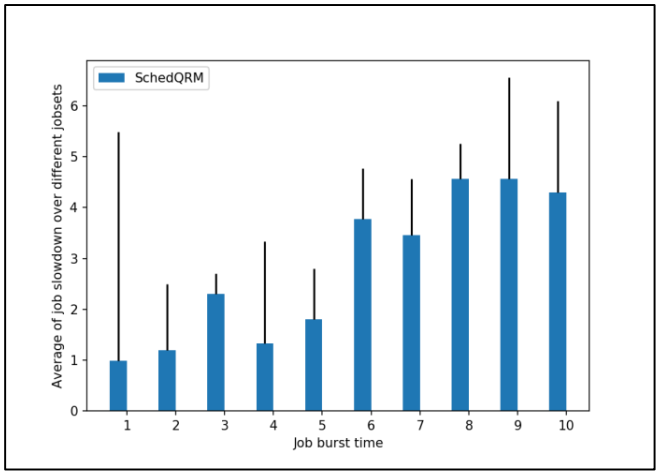


Figure 4: Performance of SchedQRM in terms of average job slowdown.

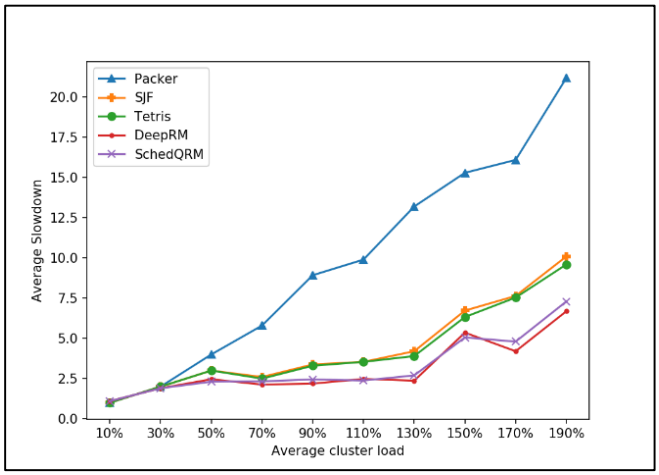


Figure 5: Job slowdown at different levels of load.

A general trend of longer jobs having higher slowdown is observed. This indicates that the *agent* is withholding longer jobs to make space for shorter jobs which helps in reducing the overall job slowdown.

Figure 5 illustrates a comparison between state-of-the-art Packer (the packing heuristic in [21]), Tetris [21], DeepRM [19] and SJF (Shortest Job First) algorithm with our proposed SchedQRM. For all cluster load values, SchedQRM performs either better or equivalent to the existing heuristics. The plot is made by averaging 100 arbitrary job sets at average cluster load value and observing the average job slowdown of all these jobs.

Conclusion

This paper presents our proposed approach for automating an end-to-end process from predicting the burst time of tasks and/or jobs until scheduling them. To achieve this, we present a 2-section model each of which performs one of the tasks stated above. Our RL *agent* focuses on the criterion of average job slowdown. The experiments show that our scheduler SchedQRM outperforms the ad-hoc heuristics. There are certain limitations faced by our model. Firstly, SchedQRM, when trained for average job slowdown, performs not as good as the DeepRM scheduler because of fixed time representation used by the authors in [19]. Our *agent* is unable to choose two jobs of the same burst time together; instead, it selects one of them and keeps the other in the *backlog*. However, such a representation makes SchedQRM much more robust and optimized, and SchedQRM is both trained and capable of working over arbitrary job sets. The second challenge is to interpret the policy used by the *agent* to reach an optimal goal. In general, it holds longer jobs to allow shorter jobs to schedule first, but interpreting the complete policy remains a challenge. We believe these challenges would further motivate research directions in the future.

References

[1] Min Z., Guodong Z., Shibing Z. and Zhihua B., “An Optimized Resource Allocation Algorithm in Cooperative Relay Cognitive Radio Networks,” 2017 Signal Processing Symposium (SPSympo).

[2] Feng T., Bi J. and Wang K., “Allocation and Scheduling of Network Resource for Multiple Control Applications in SDN,” China Communications, Volume: 12, Issue: 6, June 2015, pp. 85 – 95.

[3] Weiwei X. and Lianfeng S., “Joint Resource Allocation Using Evolutionary Algorithms In Heterogeneous Mobile Cloud Computing Networks,” China Communications, Volume: 15, Issue: 8, Aug. 2018, pp. 189 – 204.

[4] F.A. Cruz-Perez and L. Ortigoza-Guerrero, “Equal resource sharing allocation with QoS differentiation for conversational services in wireless communication networks,” IEE Proceedings - Communications, Volume: 150, Issue: 5, 14 Oct. 2003, pp. 391.

[5] Tian W., Liu L., Zhang X. and Jian C., “A Resource Allocation Algorithm Combined with Optical Power Dynamic Allocation for Indoor Hybrid VLC and Wi-Fi Network,” 2016 8th International Conference on Computational Intelligence and Communication Networks (CICN), Dec. 2016.

[6] Frances Y., Alan D. and Scott S., “A Scheduling Model for Reduced CPU Energy,” Proceedings of IEEE 36th Annual Foundations of Computer Science, Oct. 1995.

[7] Yarong C., Ling-Huey S., Ya-Chih T. Shenquan H. and Fuh-Der C., “Scheduling jobs on a single machine with dirt cleaning consideration to minimize total completion time,” IEEE Access, Volume: 7, Feb. 2019.

[8] Karim H. and Ahmed J., “A New Approach for Scheduling Tasks and/or Jobs in Big Data Cluster,” 2019 4th MEC International Conference on Big Data and Smart City (ICBDSC), Jan. 2019.

[9] Fuming L., Yunlong Z., Chaowan Y. and Xiaoyu S., “Study on the Job Shop Fuzzy Dynamic Scheduling Based on Virtual Pre-scheduling,” 2006 6th World Congress on Intelligent Control and Automation, June 2016.

SEEK DIGITAL LIBRARY

Copyright © Institute of Research Engineers and Doctors. All rights reserved.
ISBN: 978-1-63248-169-6 DOI : 10.15224/978-1-63248-169-6-09

[10] Zhao P. and Huang T., "Research of Multi-resource Dynamic Job-Shop Scheduling based on the Hybrid Genetic Algorithm," 2009 Third International Conference on Genetic and Evolutionary Computing, Oct. 2009.
[11] Xiaomin L., Jiafu W., Hong-Ning D., Muhammad I., Min X. and Antonio C., "A Hybrid Computing Solution and Resource Scheduling Strategy for Edge Computing in Smart Manufacturing," IEEE Transactions on Industrial Informatics, pp. 1-1, Feb. 2019.
[12] Liu Y., Zhong L., Zhang J. and Fu D., "Resource Isolation Policy for Task Scheduling Strategy In Open Real-Time Systems," Proceedings of the 31st Chinese Control Conference, July 2012.
[13] PhuongBang N. and Bhaskar R., "Optimal Scheduling Policies and the Performance of the CDF Scheduling," 2014 48th Asilomar Conference on Signals, Systems and Computers, Nov. 2014.
[14] S N., Mengjie Z., Mark J. and Kay T., "Automatic Design of Scheduling Policies for Dynamic Multi-objective Job Shop Scheduling via Cooperative Coevolution Genetic Programming," IEEE Transactions on Evolutionary Computation, Volume: 18, Issue: 2, April 2014, pp. 193-208.
[15] Wu Q., "Cloud Computing Task Scheduling Policy Based on Improved Particle Swarm Optimization," 2018 International Conference on Virtual Reality and Intelligent Systems (ICVRIS), Aug. 2018.
[16] Frank L., Draguna V. and Kyriakos V., "Reinforcement Learning and Feedback Control Using Natural Decision Methods to Design Optimal Adaptive Controllers," IEEE Control Systems Magazine, Volume: 32, Issue: 6, Dec. 2012, pp. 76-105.
[17] Pingan H. and S. Jagannathan, "Reinforcement Learning Neural-Network- Based Controller for Nonlinear Discrete-Time Systems With Input Constraints," IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) (Volume: 37 , Issue: 2 , April 2007), pp. 425-436.
[18] Gerald T., Nicholas J., Rajarshi D. and Mohamed B., "A Hybrid Reinforcement Learning Approach to Autonomic Resource Allocation," 2006 IEEE International Conference on Autonomic Computing, pp. 65-73.
[19] Hongzi Mao, Mohammad Alizadeh, Ishai Menachey, Srikanth Kandulay, "Resource Management with Deep Reinforcement Learning," HotNets '16 Proceedings of the 15th ACM Workshop on Hot Topics in Networks, pp. 50-56.
[20] Tom V., Spyridon S. and Branko S., "On Monte Carlo Tree Search and Reinforcement Learning", Journal of Artificial Intelligence Research, Vol 60 (2017)
[21] R. Grandl, G. Ananthanarayanan, S. Kandula, S. Rao, and A. Akella. "Multi-resource packing for cluster schedulers", SIGCOMM '14, pp. 455–466, New York, NY, USA, 2014. ACM.
[22] Diederik K. and Jimmy B., "ADAM: A Method For Stochastic Optimization", 3rd International Conference for Learning Representations, San Diego, 2015.

SEEK DIGITAL LIBRARY

Indian J Phys
https://doi.org/10.1007/s12648-019-01663-9

ORIGINAL PAPER

New holographic dark energy model with bulk viscosity in $f(R, T)$ gravity

C P Singh* and M Srivastava

Department of Applied Mathematics, Delhi Technological University, Bawana Road, New Delhi 110042, India

Received: 24 May 2019 / Accepted: 30 September 2019

Abstract: In this paper, we present the bulk viscous phenomena to mimic new holographic dark energy behavior in modified $f(R, T)$ gravity. We assume the bulk viscosity coefficient, ζ in the form of $\zeta = \zeta_0 + \zeta_1 H$, where ζ_0, ζ_1 are positive constants and H is the Hubble parameter. We obtain the exact solution of the scale factor and classify all the possible scenarios (deceleration, acceleration and their transition) of the evolution of the Universe. It is observed that the model transits from the decelerated phase to accelerated phase in early or late time depending on the values of viscous terms. For large values of viscous terms, the model always accelerates throughout the evolution. Furthermore, we also analyze two diagnostic parameters, statefinder $\{r, s\}$ and Om to discriminate our model with the other existing dark energy models. The evolutions of these diagnostics are shown in $r - s, r - q$ and Om $- z$ planes. It is found that the behavior of trajectories in different planes depends on the bulk viscous coefficients. A small combination of ζ_0 and ζ_1 gives the quintessence-like behavior, whereas a large combination gives Chaplygin gas-like model. However, the model approaches the ΛCDM model in the late time evolution of the Universe. The value of the effective equation of state parameter comes to be -0.9745 which is very close to the observational data. The entropy and generalized second law of thermodynamics are valid under certain constraints. The analysis shows that the dark energy phenomena may be explained in the presence of bulk viscosity in the modified theory of gravity.

Keywords: Dark energy cosmology; Bulk viscosity; New holographic dark energy; FRW model

PACS No.: 04.20.-q; 04.40.Nr; 04.40.Dg

1. Introduction

The observational analysis from various data [1–5] strongly suggests that the Universe is currently undergoing an accelerated expansion. A possible way of describing the accelerating Universe is provided by the introduction of a fluid with negative pressure, the so-called dark energy (DE). There are various DE models proposed in the literature, still it remains one of the open issues in cosmology. The simplest DE candidate is the cosmological constant (Λ); however, this model has some shortcomings like fine-tuning and cosmic coincidence problems.

In recent years, holographic dark energy (HDE) model has been proposed to describe DE phenomena. The concept of HDE is based on holographic principle [6, 7] which is the most important cornerstone of quantum gravity, and has great potential to solve many long-standing issues of various physical phenomena. Nowadays, it has been applied to the dark energy (DE) problem. According to the holographic principle, if the Universe is considered as a characteristic length scale L, then all the physical quantities inside the Universe, including the energy density of DE, can be described by some quantities on the boundary of the Universe. Cohen et al. [8] suggested that in quantum field theory, a short distance cutoff is related to a long distance cutoff due to the limit set by the formation of a black hole, and the total energy in a region of size L should not exceed the mass of a black hole of the same size. Under this assumption, Li [9] proposed HDE density as $\rho_\Lambda = 3c^2 M_p^2 L^{-2}$, where c^2 is a dimensionless constant and $M_p^{-2} = 8\pi G$ is the reduced Planck mass.

In the literature, many prospectives of HDE have been studied by considering various system's infrared (IR) cutoff [10–12]. Gao et al. [13] proposed another interesting

*Corresponding author, E-mail: cpsphd@rediffmail.com

Published online: 23 December 2019

© 2019 IACS

model, known as 'holographic Ricci dark energy' (HRDE) model by replacing the future event horizon with the inverse of the Ricci scalar curvature. Granda and Oliveros [14] proposed a new HDE model, known as new holographic dark energy (new HDE) model by introducing a new infrared cutoff based on purely dimensional grounds. The energy density of new HDE is given by

$$\rho_d = 3M_p^2\left(\alpha H^2 + \beta\dot{H}\right), \tag{1}$$

where α and β are constants. Wang and Xu [15] used current observations to restrict these constants. Many authors [16–19] have studied this new HDE model. Therefore, it will be interesting to work with this new HDE model.

The influence of dissipative processes including bulk viscosity, shear viscosity and heat transport plays an important role in cosmic expansion. Many authors [20–23] have investigated their role to explain the early inflation and late time evolution of the Universe. Eckart [24] was the first who formulated the relativistic thermodynamic theory of dissipative fluids which was further modified by Landau and Lifshitz [25]. Eckart theory was based on the first-order deviations from thermodynamic equilibrium. Israel and Stewart [26] developed a second-order dissipative thermodynamic theory known as the causal theory of relativistic viscosity. In this theory, the dissipative variables were included to describe non-equilibrium states due to which this theory is causal and stable. The idea of viscous DE models has been presented in recent years to understand the evolution of the Universe. In the literature [27–38], a number of works have been carried out to study the bulk viscosity as a candidate of DE, even the appropriate form of bulk viscosity can explain the accelerated expansion of the Universe without the cosmological constant [39]. These works give us the motivation to study the DE phenomena in the prospect of bulk viscosity in new HDE model.

One way to solve the cosmological problems is to modify the geometrical part of the Einstein–Hilbert action. The extension of general relativity is the $f(R)$ theories, where a function of Ricci Scalar R is used instead of only R. Nojiri and Odintsov [40], and Bamba et al. [41] have observationally confirmed the consistency of this theory. Harko et al. [42] have modified this theory and named it as $f(R, T)$ theory, where R has its usual meaning and T is the trace of energy–momentum tensor. This modified theory presents a maximal coupling between geometry and matter. In the literature [43–47], many works have been carried out to explain the different phases of the Universe in the framework of the $f(R, T)$ theory. It is thus interesting to study the new HDE model with viscosity in modified gravity theory. F. Adabi et al. [48] have studied a non-flat model filled with an interacting viscous entropy-corrected HDE with dark matter. In a paper, Karami and Fahimi [49] have discussed a non-flat FRW model filled with interacting viscous ghost dark energy and dark matter.

In a recent paper, the authors [50] have studied new HDE model with constant bulk viscosity in $f(R, T)$ gravity and have discussed the evolution of the Universe. In this paper, the most general form of bulk viscous coefficient $\zeta = \zeta_0 + \zeta_1 H$, where ζ_0 and ζ_1 are constants, has been assumed to discuss the evolution of the Universe. We discuss how the presence of viscous fluid could produce the late time acceleration. The viscous new HDE model gives time-dependent deceleration parameter which shows the phase transition. We also extend the work to discriminate the new HDE model with other existing DE models with the help of two independent geometrical diagnostics, namely statefinder pair and Om. The effective equation of state parameter, entropy and generalized second law of thermodynamics are also discussed.

The paper is organized as follows: Sect. 2 deals with field equations of viscous new HDE model in $f(R, T)$ gravity theory. Further, Sect. 2 is divided into Sects. 2.1–2.6, in which we find the solution of the scale factor, and the relevant physical and geometrical quantities such as deceleration parameter, statefinder parameters, Om diagnostic, effective equation of state parameter, entropy and generalized second law of thermodynamics, respectively, and discuss their evolutions. In Sect. 3, we present the main results and discussion. We summarize our findings in conclusion in Sect. 4. In the present paper, we assume $8\pi G = 1$ and $c = 1$.

2. Model and field equations

The action of $f(R, T)$ gravity [42] is defined as

$$S = \frac{1}{2}\int d^4x\sqrt{-g}[f(R,T) + 2\mathcal{L}_m], \tag{2}$$

where $f(R, T)$ is known as a function of the Ricci scalar R and trace T of the energy–momentum tensor $T_{\mu\nu}$ of the matter, g represents the determinant of the metric tensor $g_{\mu\nu}$, and $\mathcal{L}_m$ represents the matter Lagrangian density.

Variation of the action (2) with respect to the metric tensor $g_{\mu\nu}$ gives the field equation

$$f_R(R,T)R_{\mu\nu} - \frac{1}{2}f(R,T)g_{\mu\nu} + \left(g_{\mu\nu}\Box - \nabla_\mu\nabla_\nu\right)f_R(R,T) = T_{\mu\nu} - f_T(R,T)\left(T_{\mu\nu} + \Theta_{\mu\nu}\right), \tag{3}$$

where, the tensor $\Theta_{\mu\nu}$ is given by

$$\Theta_{\mu\nu} = -2T_{\mu\nu} + g_{\mu\nu}\mathcal{L}_m - 2g^{\alpha\beta}\frac{\partial^2\mathcal{L}_m}{\partial g^{\mu\nu}\partial g^{\alpha\beta}}. \tag{4}$$

Here, $f_R = \partial f/\partial R$, $f_T = \partial f/\partial T$, $\Box \equiv \nabla^\mu \nabla_\mu, \nabla_\mu$ is the covariant derivative.

Harko et al. [42] introduced the form of $f(R,T) = R + f(T)$, where $f(T) = \lambda T$, λ is an arbitrary constant. Using this form of $f(R, T)$, field Eq. (3) is

$$R_{\mu v} - \frac{1}{2}Rg_{\mu v} = T_{\mu v} - \lambda\left(T_{\mu v} + \Theta_{\mu v}\right) + \frac{1}{2}\lambda g_{\mu v}T. \tag{5}$$

Harko et al. [42] have suggested that the matter Lagrangian $\mathcal{L}_m$ is considered in such a way that $\mathcal{L}_m = -p$, where p stands as a thermodynamic pressure of matter of the Universe. Then, Eq. (4) becomes as $\Theta_{\mu v} = -2T_{\mu v} - pg_{\mu v}$. Now, Eq. (5) finally reduces to

$$R_{\mu v} - \frac{1}{2}Rg_{\mu v} = T_{\mu v} + \lambda\left(T_{\mu v} + pg_{\mu v}\right) + \frac{1}{2}\lambda T g_{\mu v}. \tag{6}$$

Let us consider the Friedmann–Robertson–Walker (FRW) metric in the flat space geometry ($k = 0$)

$$\mathrm{d}s^2 = \mathrm{d}t^2 - a^2(t)\left[\mathrm{d}r^2 + r^2\left(\mathrm{d}\theta^2 + \sin^2\theta \mathrm{d}\phi^2\right)\right], \tag{7}$$

and assume that the cosmic fluid possesses HDE with bulk viscosity and dust dark matter (excluding the baryonic matter). The energy–momentum tensor (EMT) can be written as

$$T_{\mu v} = (\rho_m + \rho_d)u_\mu u_v - \left(g_{\mu v} + u_\mu u_v\right)\tilde{p}_d, \tag{8}$$

where ρ_m stands for the dark matter (DM) energy density, ρ_d stands for the energy density of new HDE, and $\tilde{p}_d$ stands for an effective pressure consisting of the pressure p_d of new HDE and the bulk viscous pressure which is given by

$$\tilde{p}_d = p_d - 3\zeta H, \tag{9}$$

where ζ is the bulk viscous coefficient that arises in a fluid when it is out of local thermodynamic equilibrium. This form of effective pressure was originally proposed by Eckart [24] in the context of relativistic dissipative process occurring in thermodynamic systems went out of local thermal equilibrium.

Using line element (7) and EMT (8) into (6), we get the following equations

$$3H^2 = \rho_m + \rho_d + \lambda(\rho_m + \rho_d + p_d - 3\zeta H) + \frac{1}{2}\lambda T, \tag{10}$$

$$2\dot{H} + 3H^2 = -p_d + 3\zeta H + \frac{1}{2}\lambda T, \tag{11}$$

where $H = \dot{a}/a$ represents the Hubble parameter. An over dot denotes derivative with respect to cosmic time t. Using equation of state for NHDE $p_d = \omega_d \rho_d$, where ω_d is the equation of state parameter of new HDE, field Eqs. (10) and (11) yield

$$2\dot{H} + (1+\lambda)[\rho_m + (1+\omega_d)\rho_d] - 3(1+\lambda)\zeta H = 0. \tag{12}$$

Following the Granda and Oliveros [14], the energy density of NHDE (1) can be written as

$$\rho_d = 3\left(\alpha H^2 + \beta\dot{H}\right). \tag{13}$$

Using (8) and (11) into (10), we finally get

$$\dot{H} + (1 + 2\lambda\alpha\omega_d + \alpha\omega_d)lH^2 - (1+2\lambda)\zeta l H = 0, \tag{14}$$

where $l = \frac{3(1+\lambda)}{2+3\lambda+3\beta(1+\lambda)(1+2\lambda)\omega_d}$.

In Refs. [29, 32, 51], the most general form of bulk viscosity has been considered with generalized EoS. In this paper, we assume [29, 32, 51, 52]

$$\zeta = \zeta_0 + \zeta_1 H, \tag{15}$$

where ζ_0 and ζ_1 are two positive constants.

The term ζ is the bulk viscous coefficient that arises in a fluid when the system is out of the local thermodynamic equilibrium and which induces a viscous pressure $-\zeta\nabla_v u^v$. Of the fluid dissipative phenomena, bulk viscosity is the most favorable phenomena, compatible with the symmetry requirements of the homogeneous and isotropic FRW Universe. The parameter ζ is conventionally chosen to be positive quantity and generally depends on the cosmic time t, or redshift z, or the energy density ρ or the scale factor a. The assumption $\zeta = \zeta_0$ (constant) is the simplest parametrization for the bulk viscous coefficient. The model with the first term has been recently discussed by Srivastava and Singh [50].

In Eq. (15), the term ζ_0 takes into account the simplest parametrization for the bulk viscosity, a constant and the term $\zeta_1 H$ characterizes the possibility of a bulk viscosity proportional to the expansion rate. The motivation of considering such bulk viscous coefficient is that by fluid mechanics. We know that the transport (dissipation)/viscosity phenomenon is involved with the velocity $\dot{a}$, which is related to the scalar expansion $\theta = 3\dot{a}/a$. Therefore, a linear combination of $\zeta = \zeta_0$ (constant) and $\zeta \propto \theta$ is a more general form of bulk viscous term than considering them separately.

Using the ansatz (15) into (14), we get the evolution equation as

$$\dot{H} + lmH^2 - (1+2\lambda)\zeta_0 l H = 0, \tag{16}$$

where $m = 1 + (1+2\lambda)(\alpha\omega_d - \zeta_1)$. In what follows, we solve (16) to find the cosmological parameters, such as Hubble parameter, scale factor and deceleration parameter, and discuss the behavior in detail.

2.1. Evolution of the scale factor

On solving (16), the Hubble parameter can be obtained as

$$H(t) = H_0 e^{(1+2\lambda)l\zeta_0(t-t_0)} \left[1 + \frac{mH_0}{(1+2\lambda)\zeta_0}\left(e^{(1+2\lambda)l\zeta_0(t-t_0)} - 1\right)\right]^{-1}, \quad (17)$$

where $H = H_0$ at $t = t_0$, the present time when viscous new HDE starts to dominate. Defining $H = \dot{a}/a$, the scale factor is given by

$$a(t) = a_0\left[1 + \frac{mH_0}{(1+2\lambda)\zeta_0}\left(e^{(1+2\lambda)l\zeta_0(t-t_0)} - 1\right)\right]^{1/lm}; \quad \lambda \neq -1/2, \ \ \zeta_0 \neq 0 \quad (18)$$

where a_0 is the present scale factor at $t = t_0$. Here and thereafter, we take $a_0 = 1$. It is observed that the scale factor is of the exponential form, which can explain the phase transition. We can analyze the behavior of the scale factor for all possible combinations of (ζ_0, ζ_1) and model parameter λ. Taking $a(t) = 0$, we obtain the cosmic time when the Big Bang happens

$$t(at\ \text{Big} - \text{Bang}) = t_0 + \frac{1}{(1+2\lambda)l\zeta_0}\ln\left[1 - \frac{(1+2\lambda)\zeta_0}{mH_0}\right]. \quad (19)$$

In the early time of evolution, scale factor (18) can be approximated by

$$a \sim \left[1 + \frac{3m(1+\lambda)H_0}{2+3\lambda+3\beta(1+\lambda)(1+2\lambda)}(t-t_0)\right]^{\frac{1}{lm}}, \quad (20)$$

which varies as power-law form $a \propto t^n$ and according to which the Universe decelerates for $0 < n < 1$, has marginal inflation for $n = 1$ and accelerates for $n > 1$. Also, in the late time of evolution, the scale factor behaves as

$$a(t) \rightarrow \exp[(1+2\lambda)l\zeta_0(t-t_0)], \quad (21)$$

which behaves like de Sitter Universe. Let us compute the second-order derivatives of (18), which is given by

$$\frac{d^2a}{dt^2} = H_0^2\left(e^{2(1+2\lambda)l\zeta_0(t-t_0)}\right)\left[1 + \frac{[(1+2\lambda)\zeta_0 - mH_0]l}{H_0}e^{-(1+2\lambda)l\zeta_0(t-t_0)}\right]\left[1 + \frac{mH_0}{(1+2\lambda)\zeta_0}\left(e^{(1+2\lambda)l\zeta_0(t-t_0)} - 1\right)\right]^{\frac{1}{lm}-2}. \quad (22)$$

Equating (22) to zero, we get the transition time, t_{trans} between the decelerated and the accelerated expansion epochs, which is given by

$$t_{\text{trans}} = t_0 + \frac{1}{l(1+2\lambda)\zeta_0}\ln\left[\frac{l\{mH_0 - (1+2\lambda)\zeta_0\}}{H_0}\right]. \quad (23)$$

Using (18) into (17), the Hubble parameter in terms of the scale factor can be written as

$$H(a) = \frac{(1+2\lambda)H_0}{m}\left[\frac{\zeta_0}{H_0} + \left\{\frac{m}{(1+2\lambda)} - \frac{\zeta_0}{H_0}\right\}a^{-lm}\right]. \quad (24)$$

Differentiating (24) with respect to a, we obtain

$$\frac{d\dot{a}}{da} = \frac{(1+2\lambda)H_0}{m}\left[\frac{\zeta_0}{H_0} - \left\{\frac{m}{(1+2\lambda)} - \frac{\zeta_0}{H_0}\right\}Aa^{-lm}\right], \quad (25)$$

where $A = \left(\frac{1+3(1+\lambda)(1+2\lambda)\{(\alpha-\beta)\omega_d - \zeta_1\}}{2+3\lambda+3\beta(1+\lambda)(1+2\lambda)\omega_d}\right)$. Equating (25) to zero, the transition scale factor a_T from the decelerated to the accelerated phase can be written as

$$a_T = [A\{mH_0 - (1+2\lambda)\zeta_0\}]^{\frac{1}{lm}}, \quad (26)$$

and its transition redshift $z = a^{-1} - 1$ is given by

$$z_T = \left\{[A\{mH_0 - (1+2\lambda)\zeta_0\}]^{-\frac{1}{lm}}\right\} - 1. \quad (27)$$

From (26) and (27), we observe that the transition from decelerated to accelerated epoch takes place at the present time, i.e., at $a_T = 1$ or $z_T = 0$ for $(\zeta_0 + H_0\zeta_1) = \frac{\{1+3(1+\lambda)(1+2\lambda)(\alpha-\beta)\omega_d\}H_0}{3(1+\lambda)(1+2\lambda)}$ depending on positive or negative value of λ. In particular, substituting the observational values of model parameters $\alpha = 0.8502$ and $\beta = 0.4817$ [15] along with $\omega_d = -0.5, H0 = 1$, we get $\zeta_0 + \zeta_1 = 0.096$ for positive value of $\lambda = 0.06$ and $\zeta_0 + \zeta_1 = 0.218$ for negative value of $\lambda = -0.06$, respectively.

Figure 1 represents the evolution of the scale factor for different values of a combination of ζ_0 and ζ_1 and positive value of λ. For $0 < (\zeta_0 + \zeta_1) < 0.096$, the scale factor has a decelerated phase followed by an accelerated phase in the late time. For $(\zeta_0 + \zeta_1) = 0.096$, the transition takes place at the present time and for $(\zeta_0 + \zeta_1) > 0.096$, the transition from decelerated phase to accelerated phase occurs at early time. For higher combination of (ζ_0, ζ_1), acceleration takes

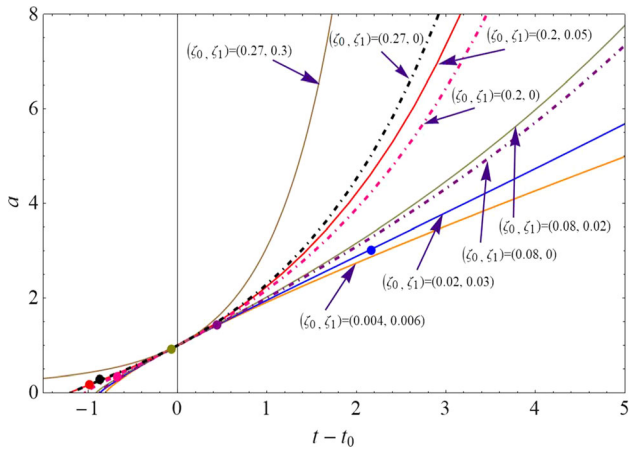


Fig. 1 The evolution of scale factor with respect to $(t-t_0)$ for different combinations of ζ_0 and ζ_1 with $\omega_d = -0.5, \lambda = 0.06, H_0 = 1, \alpha = 0.8502$ and $\beta = 0.4817$. A dot on each trajectory denotes the present time transition from the decelerated phase to accelerated phase

place in the infinite past. Similar behavior can be observed for the negative value of λ.

2.2. Deceleration parameter

Now, the evolutionary behavior of the Universe can also be discussed by the deceleration parameter (DP), which is defined by $q = -\frac{a\ddot{a}}{\dot{a}^2}$. Using (18), DP can be obtained as

$$q(t) = \left[\frac{\{mH_0 - (1+2\lambda)\zeta_0\}l}{H_0 e^{(1+2\lambda)l\zeta_0(t-t_0)}}\right] - 1. \tag{28}$$

Equation (28) shows the time-dependent value of DP which can explain the phase transition. The sign change in DP can be observed at $t = t_0$. Here, q is positive for $t < t_0$ and negative for $t > t_0$.

The DP in terms of a can be obtained as

$$q(a) = -1 + \left[\frac{lm[mH_0 - (1+2\lambda)\zeta_0]}{(1+2\lambda)\zeta_0\{a^{lm} - 1\} + mH_0}\right]. \tag{29}$$

Using the relation $a = (1+z)^{-1}$, we can also write the expression of DP in terms of redshift z. Now, the present value of DP corresponding to $z = 0$ or $a = 1$ is,

$$q_0 = \frac{[mH_0 - (1+2\lambda)\zeta_0]l}{H_0} - 1. \tag{30}$$

This equation shows that if $(\zeta_0 + H_0\zeta_1) = \frac{\{1+3(1+\lambda)(1+2\lambda)(\alpha-\beta)\omega_d\}H_0}{3(1+\lambda)(1+2\lambda)}$, the value of $q_0 = 0$. Thus, the transition into accelerating phase would occur at the present time for this combination of (ζ_0, ζ_1). In particular, taking $\alpha = 0.8502, \beta = 0.4817, \omega_d = -0.5, H_0 = 1$ and $\lambda = 0.06$ in the above expression, we get $\zeta_0 + \zeta_1 = 0.096$, which gives $q_0 = 0$. We plot the trajectories of q with a for different combinations of (ζ_0, ζ_1) as shown in Fig. 2. We can also observe the difference with constant bulk viscosity. A very fast transition can be observed with (ζ_0, ζ_1).

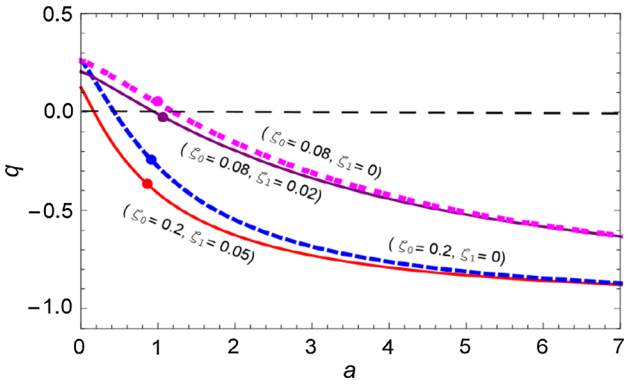


Fig. 2 The evolution of deceleration parameter with respect to a for $\omega_d = -0.5, \lambda = 0.06, H_0 = 1, \alpha = 0.8502$ and $\beta = 0.4817$. A dot on each trajectory denotes the present value q_0

2.3. Statefinder diagnostic

Sahni et al. [53] and Alam et al. [54] proposed a new geometrical diagnostic pair $\{r, s\}$ called statefinder, which is constructed from a space–time metric directly, and allows us to characterize the properties of DE in a model-independent manner. It is dimensionless and is constructed from the scale factor $a(t)$ and its derivatives up to the third order. The statefinder pair is defined as

$$r = \frac{\dddot{a}}{aH^3} \quad \text{and} \quad s = \frac{r - 1}{3(q - 1/2)}. \tag{31}$$

Using the values of scale factor, Hubble parameter and deceleration parameter into (31), we get

$$r = 1 + \left[\frac{l(3-lm)\{(1+2\lambda)\zeta_0 - mH_0\}}{H_0 e^{(1+2\lambda)l\zeta_0(t-t_0)}}\right] \left[\frac{l^2\{(1+2\lambda)\zeta_0 - mH_0\}^2}{H_0^2 e^{2(1+2\lambda)l\zeta_0(t-t_0)}}\right]. \tag{32}$$

and

$$s = \frac{\frac{2l\left(1-\frac{lm}{3}\right)\{(1+2\lambda)\zeta_0 - mH_0\}}{3H_0 e^{(1+2\lambda)l\zeta_0(t-t_0)}} + \frac{2l^2\{(1+2\lambda)\zeta_0 - mH_0\}^2}{9H_0^2 e^{2(1+2\lambda)l\zeta_0(t-t_0)}}}{\frac{2l\{mH_0 - (1+2\lambda)\zeta_0\}}{3H_0 e^{(1+2\lambda)l\zeta_0(t-t_0)}} - 1}, \tag{33}$$

which are time-dependent due to the presence of bulk viscosity coefficient. Equations (32) and (33) show that as $(t - t_0) \to \infty, \{r, s\} \to \{1, 0\}$, a value corresponding to the Lambda cold dark matter (ΛCDM) model. Hence, the viscous new HDE model resembles the ΛCDM model in future. The $r - s$ and $r - q$ trajectories for different combinations of (ζ_0, ζ_1) and positive value of λ (e.g., $\lambda = 0.06$) are shown in Figs. 3 and 4, respectively. Figures 5 and 6 show the respective $r - s$ and $r - q$ trajectories for different combinations of (ζ_0, ζ_1) and negative value of λ (e.g., $\lambda = -0.06$).

In Figs. 3 and 5, stars represent the fixed point values of ΛCDM and SCDM models, and a dot on each trajectory represents the present time values of $\{r, s\} = \{r_0, s_0\}$ and $\{r, q\} = \{r_0, q_0\}$, where r_0, s_0 and q_0 represent the present time values of r, s and q, respectively. The $r - s$ planes in Figs. 3 and 5 are divided into two regions $r < 1, s > 0$ and $r > 1, s < 0$ by a vertical line passing through the point (1,0).

The trajectories in the $\{r, s\}$ planes lying in the region $r < 1, s > 0$, a feature similar to the quintessence model (Q-model) of DE [53, 54]. The trajectories in the $\{r, s\}$ planes lying in the region $r > 1, s < 0$, a feature similar to the generalized Chaplygin gas (CG) model of DE [55]. Here, we obtain a parabolic trajectory for both cases on λ. From Fig. 3, for $\lambda = 0.06$ we notice that the model behaves like Q-model for $0 < (\zeta_0 + \zeta_1) \leq 0.46$, whereas for

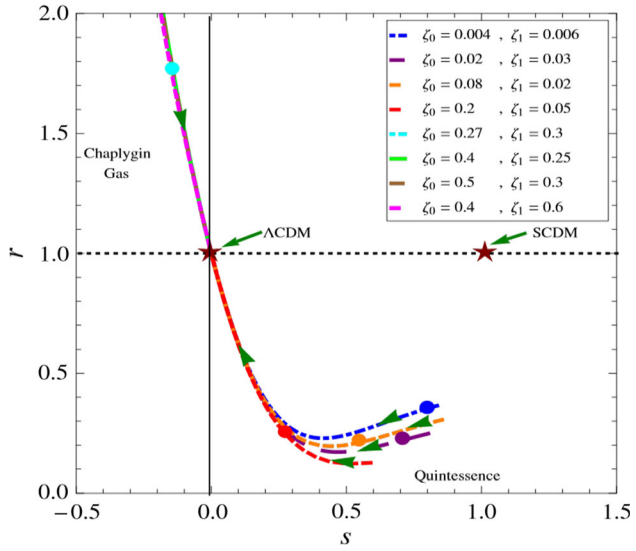


Fig. 3 The $r-s$ trajectories in $r-s$ plane are plotted for different combinations of (ζ_0, ζ_1) with $\omega_d = -0.5, \alpha = 0.8502, \beta = 0.4817$ and $\lambda = 0.06$. The arrows represent the direction of the evolutions of statefinder diagnostic pair with time

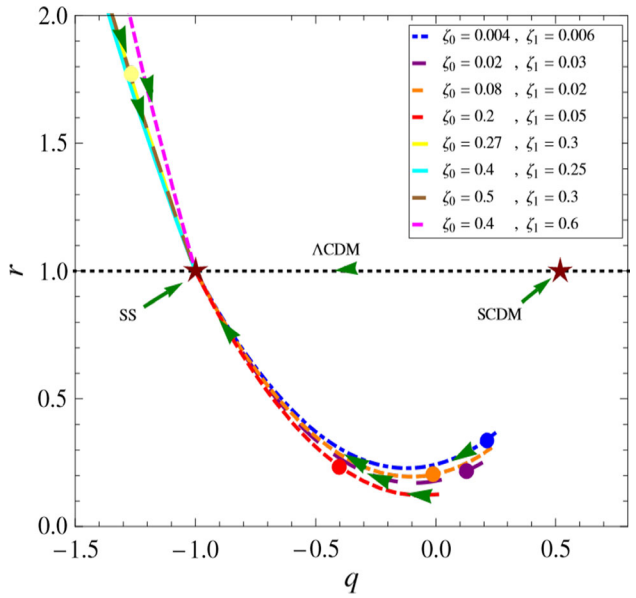


Fig. 4 The $r-q$ trajectories in $r-q$ plane are plotted for different combinations of (ζ_0, ζ_1) with $\omega_d = -0.5$, $\alpha = 0.8502, \beta = 0.4817$ and $\lambda = 0.06$. The arrows represent the direction of the evolutions of statefinder diagnostic pair with time

$(\zeta_0 + \zeta_1) > 0.46$ the model mimics like CG model. The trajectories in both the regions converge to ΛCDM model in late time of evolution.

In the case $\lambda = -0.06$, we observe from Fig. 5 that all the $r-s$ trajectories lie in the region $r<1, s>0$ for $0<(\zeta_0+\zeta_1)\leq 0.71$ which imply that the viscous new HDE model corresponds to Q model while the trajectories lie in the $(r>1, s<0)$ region for $(\zeta_0+\zeta_1) > 0.71$, i.e., the

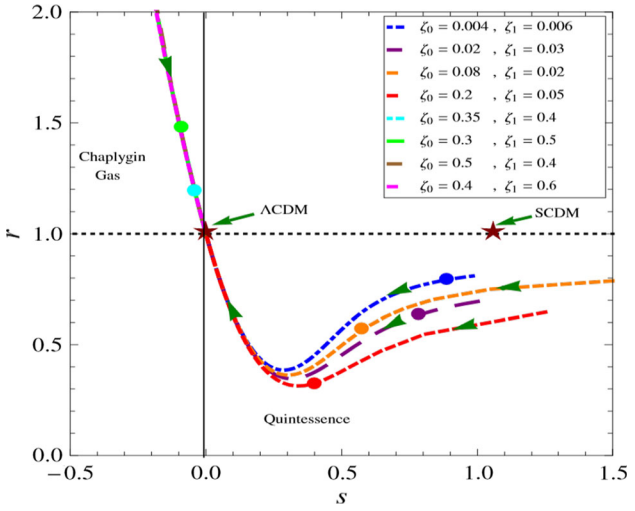


Fig. 5 The $r-s$ trajectories in $r-s$ plane are plotted for different combinations of (ζ_0, ζ_1) with $\omega_d = -0.5, \alpha = 0.8502, \beta = 0.4817$ and $\lambda = -0.06$. The arrows represent the direction of the evolutions of statefinder diagnostic pair with time

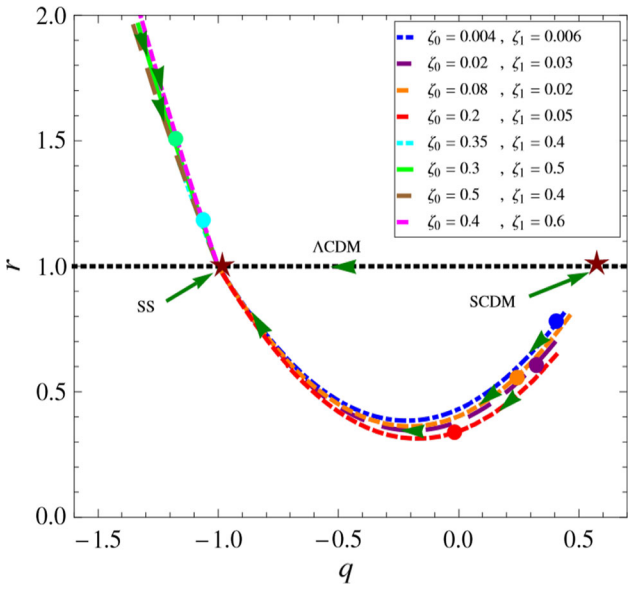


Fig. 6 The $r-q$ trajectories in $r-q$ plane are plotted for different combinations of (ζ_0, ζ_1) with $\omega_d = -0.5, \alpha = 0.8502, \beta = 0.4817$ and $\lambda = -0.06$. The arrows represent the direction of the evolutions of statefinder diagnostic pair with time

model mimics like the *CG* model. In the late time, the viscous new HDE model approaches the ΛCDM. For some combinations like $(\zeta_0, \zeta_1) = (0.08, 0.02)$, the trajectory starts in the neighborhood of the *SCDM* model and approaches the ΛCDM. Thus, we can conclude that for any value of λ (either positive or negative), our viscous new HDE model in the framework of $f(R,T)$ theory mimics like Q and *CG* models for specific range of viscosity coefficients, and in the late time of evolution it always converges to the ΛCDM model.

The $r-q$ trajectories in $r-q$ plane for positive and negative values of λ and for different combinations of (ζ_0, ζ_1) are shown in Figs. 4 and 6, respectively. Here, in both the figures stars represent the fixed point values $\{r, q\} = \{1, 0.5\}$ for SCDM model and $\{r, q\} = \{1, -1\}$ for steady-state (*SS*) model. In figures, arrows are used to show the direction of the trajectories and the time evolution of the ΛCDM model is represented by the horizontal line at $r = 1$. It can be observed that q changes its sign from positive to negative with respect to time for $0 < (\zeta_0 + \zeta_1) \leq 0.46$ in case of $\lambda = 0.06$ and for $0 < (\zeta_0 + \zeta_1) \leq 0.71$ in case $\lambda = -0.06$, which show the Universe transits from decelerated to accelerated phase. For $(\zeta_0 + \zeta_1) > 0.46$ when $\lambda = 0.06$ and $(\zeta_0 + \zeta_1) > 0.71$ when $\lambda = -0.06$, q is always negative showing the behavior of phantom. In the beginning, this new HDE model acts different from the ΛCDM model but in the late time it behaves the same as the ΛCDM which converges to *SS* model in the late time evolution.

From Fig. 6, we analyze that for some small values of (ζ_0, ζ_1), like $(\zeta_0, \zeta_1) = (0.08, 0.02)$, the $r-q$ trajectory starts from the neighborhood of the *SCDM* model. The present position of $\{r, s\} = \{r_0, s_0\}$ is indicated by dot in the plot. This means that the present viscous new HDE model is distinguishably different from the ΛCDM model but in the late time it converges to *SS* model.

The present viscous new HDE model can also be discriminated from the holographic dark energy model with event horizon as the infrared cutoff, in which the $r-s$ evolution starts from a region $r \sim 1$, $s \sim 2/3$ and ends on the ΛCDM point [56]. It can also be discriminated from Ricci dark energy model, in which $r-s$ trajectory is a vertical segment, i.e., s is a constant during the evolution of the Universe [57].

2.4. Om diagnostic

Sahni et al. [58] introduced a new geometrical diagnostic known as Om, which is defined as

$$\mathrm{Om}(z) = \frac{\frac{H^2(z)}{H_0^2} - 1}{(1+z)^3 - 1}. \tag{34}$$

Substituting the value of $H(z)$ from (24) into (34), the $\mathrm{Om}(z)$ is obtained as

$$\mathrm{Om}(z) = \frac{\frac{(1+2\lambda)^2}{m^2}\left[\frac{\zeta_0}{H_0} + \left\{\frac{m}{(1+2\lambda)} - \frac{\zeta_0}{H_0}\right\}(1+z)^{lm}\right]^2 - 1}{\left[(1+z)^3 - 1\right]}. \tag{35}$$

The trajectory in Fig. 7 is divided horizontally into two regions. In the lower region, it may be seen that $\mathrm{Om}(z)$ decreases as z decreases for $(\zeta_0 + \zeta_1) > 0.46$, so positive slope of $\mathrm{Om}(z)$ suggests phantom-like $(\omega_d < -1)$ behavior in the presence of viscosity with positive values of $\lambda = 0.06$. However, in the upper region, $\mathrm{Om}(z)$ increases as z decreases for $0 < (\zeta_0 + \zeta_1) \leq 0.46$, so negative slope of $\mathrm{Om}(z)$ indicating quintessence-like behavior in the presence of viscosity.

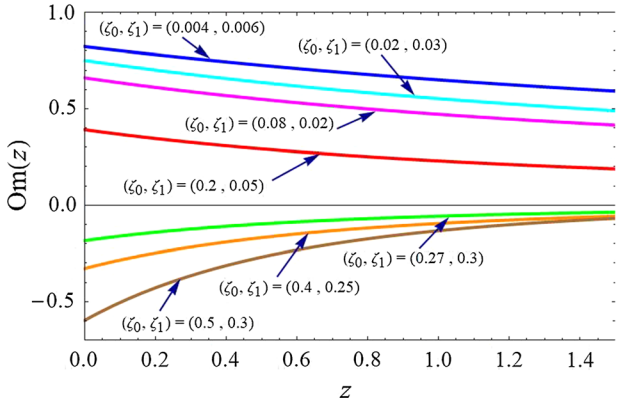


Fig. 7 $\mathrm{Om}(z) - z$ trajectories for different values of ζ_0 and ζ_1 with $\omega_d = -0.5$, $\lambda = 0.06$, $\alpha = 0.8502$ and $\beta = 0.4817$

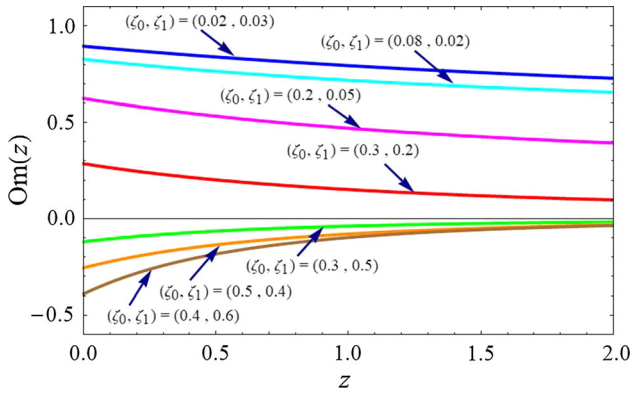


Fig. 8 $\mathrm{Om} - z$ trajectories for different values of ζ_0 and ζ_1 with $\omega_d = -0.5$, $\lambda = -0.06$, $\alpha = 0.8502$ and $\beta = 0.4817$

Similarly, the trajectory in Fig. 8 is divided horizontally into two regions. In the lower region, it may be seen that $\mathrm{Om}(z)$ decreases as z decreases for $(\zeta_0 + \zeta_1) > 0.71$, so positive slope of $\mathrm{Om}(z)$ suggests phantom-like $(\omega_d < -1)$ behavior in the presence of viscosity with negative values of $\lambda = -0.06$. However, in the upper region, $\mathrm{Om}(z)$ increases as z decreases for $0 < (\zeta_0 + \zeta_1) \leq 0.71$, so negative slope of $\mathrm{Om}(z)$ indicating quintessence-like behavior in the presence of viscosity.

In the late time of evolution when $z = -1$, we get $\mathrm{Om}(z) = 1 - \frac{(1+2\lambda)^2\zeta_0^2}{[1+(1+2\lambda)(\alpha\omega_d - \zeta_1)]^2 H_0^2}$, which is the constant value of $\mathrm{Om}(z)$, i.e., zero curvature. Thus, in the late time the viscous new HDE corresponds to the ΛCDM.

2.5. Equation of state parameter

Let us discuss the bulk viscous effect on effective equation of state parameter, ω_{eff} which is given by

$$\omega_{\text{eff}} = \frac{p_d - 3\zeta H}{\rho_m + \rho_d}, \tag{36}$$

where $p_m = 0$. On substituting the values of p_d, ζ, H, ρ_m and ρ_d in Eq. (36), we get

$$\omega_{\text{eff}} = \frac{(2+3\lambda)\left[\frac{\{(1+2\lambda)l\beta\omega_d - 1\}[(1+2\lambda)\zeta_0 - mH_0]}{e^{(1+2\lambda)l\zeta_0(t-t_0)}} - H_0\right]}{\left[\frac{\{(1+2\lambda)l\beta\omega_d - 1\}[(1+2\lambda)\zeta_0 - mH_0]\lambda}{e^{(1+2\lambda)l\zeta_0(t-t_0)}} + \left(2+3\lambda+\alpha\lambda+2\alpha\lambda^2\right)H_0\right]}. \tag{37}$$

The evolutions of ω_{eff} versus t are shown in Figs. 9 and 10 for different pairs of (ζ_0, ζ_1) in respect of $\lambda = 0.06$ and $\lambda = -0.06$, respectively. Figure 9 shows that the trajectories of ω_{eff} start from $\omega_{\text{eff}} > -1$ (it may also start from matter-dominated era) for small values and $\omega_{\text{eff}} < -1$ for large values of (ζ_0, ζ_1) in respect of $\lambda = 0.06$. As $t \to \infty$, ω_{eff} approaches to a constant for all values of (ζ_0, ζ_1), i.e., $\omega_{\text{eff}} \to -0.9745$. There is no $\omega_{\text{eff}} = -1$ crossing for small values of (ζ_0, ζ_1), but for large values of (ζ_0, ζ_1) it will cross the $\omega_{\text{eff}} = -1$. In Fig. 10, where we have $\lambda = -0.06$, we can observe the similar evolution of ω_{eff}. However, $\omega_{\text{eff}} \to -1.0252$ in the late time, i.e., it crosses $\omega_{\text{eff}} = -1$ for small values of (ζ_0, ζ_1) but it will not cross for large values of (ζ_0, ζ_1). Thus, the effective EoS parameter for both models shows consistency with the observational data given in Ref. [59]. We can say that the dark energy phenomena may be obtained in the presence of viscous fluid.

2.6. Entropy and second law of thermodynamics

The local entropy production for a fluid on a flat FRW space-time is expressed as [60]

$$T\nabla_\mu s^\nu = \zeta(\nabla_\nu u^\nu)^2 = 9H^2\zeta, \tag{38}$$

where T is the temperature, $\nabla_\nu s^\nu$ is the rate at which entropy is being generated in unit volume.

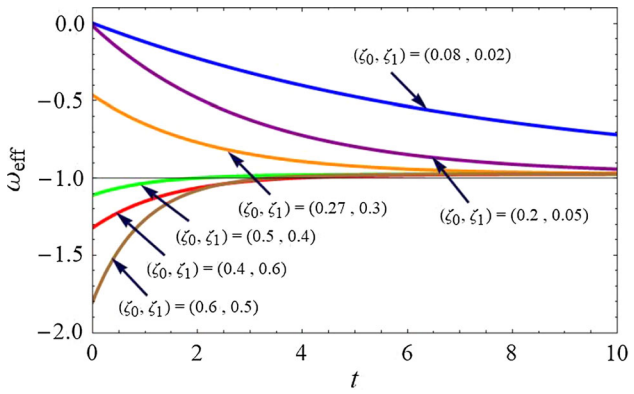


Fig. 9 The evolution of ω_{eff} for different values of (ζ_0, ζ_1) in respect of $\lambda = 0.06$. We take $H_0 = 1, \omega_d = -0.5, \alpha = 0.8502$ and $\beta = 0.4817$

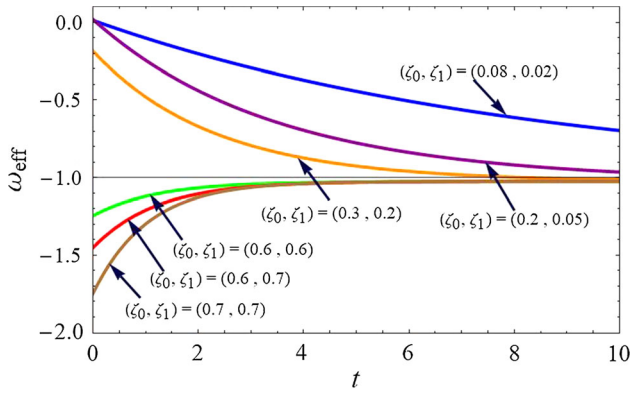


Fig. 10 The evolution of ω_{eff} for different values of (ζ_0, ζ_1) in respect of $\lambda = -0.06$. We take $H_0 = 1, \omega_d = -0.5, \alpha = 0.8502$ and $\beta = 0.4817$

The second law of thermodynamics can be stated as

$$T\nabla_\mu s^\nu \geq 0. \tag{39}$$

Since the Hubble parameter H is positive in an expanding Universe, ζ has to be positive in order to preserve the validity of the second law of thermodynamics. Thus, for the present model Eq. (38) implies the following inequality

$$\zeta = \zeta_0 + \zeta_1 H \geq 0. \tag{40}$$

Using (24), we find the expression for the total bulk viscosity $\zeta(a)$ as

$$\zeta(a) = \zeta_0 + \zeta_1\left[\frac{(1+2\lambda)H_0}{m}\left(\frac{\zeta_0}{H_0} + \left\{\frac{m}{(1+2\lambda)} - \frac{\zeta_0}{H_0}\right\}a^{-lm}\right)\right]. \tag{41}$$

The value of the scale factor, at which the transition of the total bulk viscosity between negative and positive values happens, is

$$a_{\text{np}} = \left[\frac{\{1+(1+2\lambda)\alpha\omega_d\}\zeta_0}{\{(1+2\lambda)(\zeta_0 + \zeta_1 H_0 - \alpha\omega_d) - 1\}\zeta_1}\right]^{-\frac{1}{lm}}, \tag{42}$$

where the subscript "np" stands for "negative to positive" values. In the late time of evolution, i.e., at $a \to \infty$, the total bulk viscosity is $\zeta(a) = \frac{\{1+(1+2\lambda)\alpha\omega_d\}\zeta_0}{m}$.

Figures 11 and 12 represent the evolution of the total bulk viscous coefficient ζ with respect to the scale factor for the different combinations of the (ζ_0, ζ_1) with positive $(\lambda = 0.06)$ and negative $(\lambda = -0.06)$ values of the coupling parameter, respectively. Figures 10 and 11 show that the total bulk viscosity is positive throughout the evolution of the Universe for $0 < (\zeta_0 + \zeta_1) \leq 0.46$ with $\lambda = 0.06$, and $0 < (\zeta_0 + \zeta_1) \leq 0.71$ with $\lambda = -0.06$, respectively. These curves have been shown above the line $(0, 0)$. Therefore, the model does not violate the entropy law for these ranges.

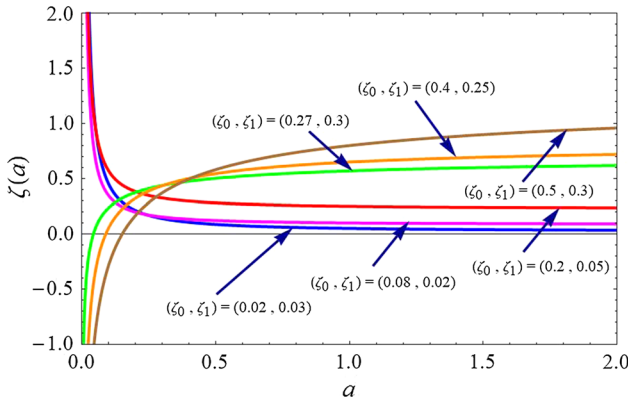


Fig. 11 The evolution of $\zeta(a)$ for different combinations of ζ_0 and ζ_1 with $\omega_d = -0.5, \lambda = 0.06, \alpha = 0.8502$ and $\beta = 0.4817$

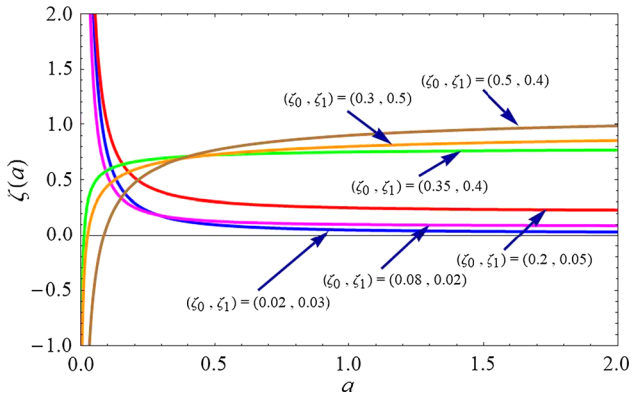


Fig. 12 The evolution of $\zeta(a)$ for different combinations of ζ_0 and ζ_1 with $\omega_d = -0.5, \lambda = -0.06, \alpha = 0.8502$ and $\beta = 0.4817$

The figures also show that the total bulk, viscous coefficient is evolving from negative to positive value for $(\zeta_0 + \zeta_1) \geq 0.46$ and $(\zeta_0 + \zeta_1) \geq 0.71$, respectively. Thus, the rate of entropy production is negative for large values of $(\zeta_0 + \zeta_1)$ in early epoch and positive in the later epoch. Hence, the entropy law violates in the early epoch and obeys in the later epoch for these values of $(\zeta_0 + \zeta_1)$.

In an absolute way, the status of the second law of thermodynamics should be considered along with the accounting of the entropy generation from the horizon. In those circumstances, the second law becomes the generalized second law (GSL) of thermodynamics, according to which the total sum of the entropies of the fluid components of the Universe plus that of the horizon entropy should never decrease [61, 62]. In the present model, this means the rate of change in complete entropy along with horizon must be greater than or equal to 0. Mathew et al. [62] have discussed the status of GSL for flat FRW Universe with matter and cosmological vacuum. Karami et al. [63] have discussed the status of GSL in a flat Universe with viscous dark energy and have shown that the GSL is valid in FRW Universe with apparent horizon as the boundary. Many authors [64, 65] have verified the status of GSL with apparent horizon as the boundary. The status of GSL in the modified gravity theories has also been discussed by many authors [43, 66–68] with the apparent horizon as the boundary.

Let us verify the GSL of thermodynamics for this model. As stated above, we consider the apparent horizon as the boundary of the Universe. Then, the GSL can be stated as

$$\frac{d}{dt}(S_{\text{tot}}) = \frac{d}{dt}\left(S_m + S_d + S_p + \tilde{S}_h\right) \geq 0, \tag{43}$$

where S_m, S_d, S_p and $\tilde{S}_h$ are the entropies of the dark matter, dark energy, the entropy production and the entropy of the apparent horizon, respectively.

From the first law of thermodynamics, the change in entropy of the viscous matter inside the apparent horizon can be obtained using Gibbs equation

$$T_i dS_i = d(\rho_i V) + p_i dV, \tag{44}$$

where T_i is the temperature and $S_i = S_m + S_d$ is the sum of the entropies of the dark matter and dark energy, ρ_i represents the sum of the densities of the dark matter and dark energy, $V = \frac{4\pi\tilde{r}_h^3}{3}$ is the volume of the apparent horizon with $\tilde{r}_h$ as a radius of the apparent horizon and $p_i = \tilde{p}_d = p_d - 3\zeta H$ as the effective pressure.

In the present viscous model, the above Gibbs equation modifies to [43, 67, 68]

$$T_i dS_i = d(\rho_i V) + p_i dV - T_i dS_p. \tag{45}$$

The radius $\tilde{r}_h$ of the apparent horizon for a flat FRW Universe is defined as [68]

$$\tilde{r}_h = \frac{1}{H}. \tag{46}$$

In the $f(R, T)$ theories, the entropy associated with the apparent horizon is defined as

$$\tilde{S}_h = \frac{\tilde{A}_h}{4}\frac{f_R(R,T)}{\left(\frac{1}{8\pi} + \frac{f_T(R,T)}{8\pi}\right)} = \frac{2\pi\tilde{A}_h}{(1+\lambda)} = \frac{8\pi^2\tilde{r}_h^2}{(1+\lambda)}, \tag{47}$$

where $\tilde{A}_h = 4\pi\tilde{r}_h^2$ is the area of the apparent horizon. In this paper, we have taken $8\pi G = 1$. Taking the derivative of Eq. (47) with respect to time t, we get

$$\dot{\tilde{S}}_h = \frac{16\pi^2\tilde{r}_h\,\dot{\tilde{r}}_h}{(1+\lambda)} = \frac{16\pi^2}{(1+\lambda)H}\left(-\frac{\dot{H}}{H^2}\right). \tag{48}$$

Now, from (45) we have

$$\begin{aligned} T_i\left(\dot{S}_m + \dot{S}_d + \dot{S}_p\right) &= 4\pi\tilde{r}_h^2\left(\dot{\tilde{r}}_h - 1\right) \\ &\left\{\rho_m + (1+\omega_d)\rho_d - 3\zeta_0 H - 3\zeta_1 H^2\right\}. \end{aligned} \tag{49}$$

Under the thermal equilibrium conditions between the fluids and the horizon, we have $T_i = \tilde{T}_h$. We take the

temperature $\tilde{T}_h = \frac{H}{2\pi}$, which is equal to the Hawking temperature of the horizon with the assumption that the fluid within the horizon is in equilibrium with the horizon, so there is no effective flow of the fluid toward the horizon [69]. Now, Eq. (49) becomes as

$$\dot{S}_m + \dot{S}_d + \dot{S}_p = \frac{8\pi^2}{H^3}\left(-\frac{\dot{H}}{H^2} - 1\right) \{\rho_m + (1+\omega_d)\rho_d - 3\zeta_0 H - 3\zeta_1 H^2\}. \tag{50}$$

Now, using the definition of deceleration parameter $q = \left(-\frac{\dot{H}}{H^2} - 1\right)$ and Eqs. (48) and (50) into (43), we get the change in the sum of all the entropies as

$$\begin{aligned}\dot{S}_{tot} &= \dot{S}_m + \dot{S}_d + \dot{S}_p + \dot{\tilde{S}}_h \\ &= \frac{8\pi^2 q}{H^3}\{\rho_m + (1+\omega_d)\rho_d - 3\zeta_0 H - 3\zeta_1 H^2\} \\ &\quad + \frac{16\pi^2(q+1)}{(1+\lambda)H}.\end{aligned} \tag{51}$$

On substituting the required values in above equation and simplifying, we get the change in total entropy as

$$\begin{aligned}\dot{S}_{\rm tot} = &\left[\frac{16\pi^2[mH_0 - (1+2\lambda)\zeta_0]\left\{1 + \frac{mH_0\left(e^{(1+2\lambda)l\zeta_0(t-t_0)} - 1\right)}{(1+2\lambda)\zeta_0}\right\}}{H_0^2 e^{2(1+2\lambda)l\zeta_0(t-t_0)}}\right] \\ &\left[\frac{l}{(1+\lambda)} + \frac{3\{1-(1+2\lambda)l\beta\omega_d\}}{(2+3\lambda)H_0}\right. \\ &\left.\left\{\frac{[mH_0 - (1+2\lambda)\zeta_0]l}{e^{(1+2\lambda)l\zeta_0(t-t_0)}} - H_0\right\}\right].\end{aligned} \tag{52}$$

From Eq. (52), it is noticed that the rate of change in entropy cannot clearly state either the change in entropy is greater than and equal to zero. We plot the evolution of $\dot{S}$ in Figs. 13 and 14, respectively, for positive and negative values of the λ (e.g., $\lambda = 0.06, \lambda = -0.06$) and different combinations of viscosity coefficients (ζ_0, ζ_1). Figure 13 shows that the GSL is always valid for $\zeta_0 > 0$ and

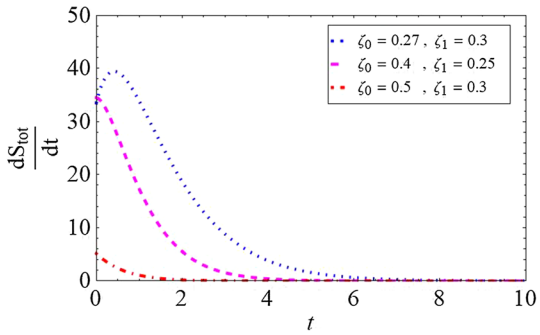


Fig. 13 The change in total entropy with respect to t for different combinations of ζ_0 and ζ_1 with $\omega_d = -0.5, \lambda = 0.06, \alpha = 0.8502$ and $\beta = 0.4817$

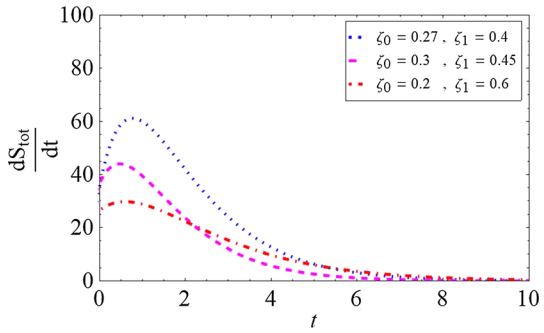


Fig. 14 The change in total entropy with respect to t for different combinations of ζ_0 and ζ_1 with $\omega_d = -0.5, \lambda = -0.06, \alpha = 0.8502$ and $\beta = 0.4817$

$0.27 \le \zeta_1 \le 0.85$, whereas it holds for $\zeta_0 > 0$ and $0.39 \le \zeta_1 \le 1$ in Fig. 14. We also notice that in both the cases, the total entropy corresponds to zero in the late time of the evolution.

3. Results and Discussion

The field equations of viscous NHDE model in modified $f(R, T)$ gravity have been solved for the most general bulk viscous coefficient. The scale factor has been obtained in the form of exponential which describes the phase transition of the Universe. We have analyzed the behavior of the scale factor for all possible combinations of (ζ_0, ζ_1) and model parameter λ. We have obtained the cosmic time when the Big Bang happens. In the early time, the scale factor is approximated by power-law expansion which shows that the model expands with decelerated rate in the early time. In the late time, the scale factor behaves as a de Sitter-like expansion, i.e., expands with accelerated rate. This shows that the scale factor at the respective limits has an earlier decelerated phase followed by an accelerated phase in the later stage of the evolution. We have calculated the transition time between the decelerated and accelerated epoch. We have expressed Hubble parameter in terms of scale factor to get the transition between the decelerated and accelerated epoch in terms of scale factor a_T and redshift z_T. Thus, for the present time transition from deceleration to acceleration taking place at $\zeta_0 + \zeta_1 = 0.096$ for positive values of $\lambda = 0.06$ and for negative value of $\lambda = -0.06$, we get $\zeta_0 + \zeta_1 = 0.218$. A plot of the evolution of the scale factor is given in Fig. 1. For different values of a combination of ζ_0 and ζ_1 and positive value of λ. For $0 < (\zeta_0 + \zeta_1) < 0.096$, the scale factor has an earlier deceleration phase followed by an acceleration phase in the late time. For $(\zeta_0 + \zeta_1) > 0.096$, the transition takes place in the past of the Universe. Similar behavior can be observed for the negative value of λ.

The assumption $\zeta = \zeta_0$ (constant) is the simplest parametrization for the bulk viscous coefficient. In this

paper, we have considered a more general form of bulk viscosity which is time-dependent and have obtained the solutions accordingly. In the absence of the second term, i.e., $\zeta_1 = 0$, the solution reduces to [50] which shows that this work is generalization of the previous work. From Fig. 1, it is observed that the transition from a decelerating to an accelerating regime occurs at a very early time (transition time) due to the inclusion of a second term related to the expansion rate. Thus, the addition of the second term in bulk viscous coefficient increases the rate of expansion and decreases the transition time.

We have obtained time-dependent DP which shows transition from one phase to other. As $(t-t_0) \to \infty$, we get $q \to -1$, which shows that the model accelerates in the late time. We have observed that for $\zeta_0 + \zeta_1 = 0.096$, we have $q_0 = 0$. Therefore, the transition into acceleration phase would occur at the present time. It differs in the current value of deceleration parameter, q_0, e.g., for $(\zeta_0, \zeta_1) = (0.2, 0.05)$, $q_0 = -0.413$ and for $(\zeta_0, \zeta_1) = (0.2, 0)$, $q_0 = -0.278$. So, it gives better value in the presence of both of the coefficients. The corresponding trajectory of q for $\zeta = \zeta_0 + \zeta_1 H$ and $\zeta = \zeta_0$ is shown in Fig. 2. It is found that a very early transition occurs with variable coefficient of bulk viscosity. The fast rate of transition can be observed in this case.

We have discussed two geometrical diagnostics: statefinder and Om, to discriminate with the other existing DE models. Let us first discuss about the statefinder diagnostic for viscous new HDE model. In the late time as $(t-t_0) \to \infty$, the statefinder pair $\{r, s\}$ for the viscous new HDE model tends to $\{1, 0\}$, a value corresponding to the ΛCDM model. We have plotted the trajectory of statefinder pair $\{r, s\}$ in $r - s$ plane for different combinations of (ζ_0, ζ_1) as shown in Figs. 3 and 5 for positive and negative values of coupling parameter λ, respectively. It has been observed that for any value of λ (either positive or negative), our viscous new HDE model in the framework of $f(R, T)$ theory mimics like Q and CG models for specific range of viscosity coefficients, and in the late time of evolution it always converges to the ΛCDM model.

We have also plotted the trajectory of $\{r, q\}$ for different combinations of (ζ_0, ζ_1) and positive and negative values of λ as shown in Figs. 4 and 6. It can be observed that q changes its sign from positive to negative with respect to time for $0 < (\zeta_0 + \zeta_1) \leq 0.46$ in case of $\lambda = 0.06$ and for $0 < (\zeta_0 + \zeta_1) \leq 0.71$ in case of $\lambda = -0.06$, which show the phase transition from decelerated phase to accelerated phase. For $(\zeta_0 + \zeta_1) > 0.46$ when $\lambda = 0.06$ and $(\zeta_0 + \zeta_1) > 0.71$ when $\lambda = -0.06$, q is always negative showing the behavior of phantom. Thus, in the beginning this model behaves different from the ΛCDM model but in the late time it behaves the same as the ΛCDM which converges to SS model in the late time evolution. The present viscous new HDE model can also be discriminated from the holographic dark energy model with event horizon as the infrared cutoff, in which the $r - s$ evolution starts from a region $r \sim 1, s \sim 2/3$ and ends on the ΛCDM point [56]. It can also be discriminated from Ricci dark energy model in which the $r - s$ trajectory is a vertical segment, i.e., s is a constant during the evolution of the Universe [57].

The second geometrical diagnostic, namely Om, has been carried out in Sect. 2.4. We have plotted the Om(z) versus redshift z trajectories for different combinations of (ζ_0, ζ_1) in respect of positive and negative values of λ as shown in Figs. 7 and 8, respectively. We have observed two types of trajectory, one has the positive curvature which suggests the phantom-like behavior and the second has the negative curvature which suggests the quintessence-like behavior.

We have also calculated the effective EoS parameter and analyzed the evolution of it graphically for different suitable values of viscosity coefficients. We have observed that $\omega_{\text{eff}} \to -0.9745$ and -1.0252, respectively, as $t \to \infty$ in case of positive and negative values of λ which is very close to -0.93 predicted in Ref. [70].

At the end, we have discussed the entropy and second law of thermodynamics for viscous new HDE model. Figures 11 and 12 plot the evolution of the total bulk viscous coefficient. It has been observed that the total bulk viscosity is positive throughout the evolution of the Universe for $0 < (\zeta_0 + \zeta_1) \leq 0.46$ with $\lambda = 0.06$, and $0 < (\zeta_0 + \zeta_1) \leq 0.71$ with $\lambda = -0.06$, respectively. Therefore, the model does not violate the entropy law for these ranges. The figures also show that $\zeta(a)$ is evolving from negative to positive value for $(\zeta_0 + \zeta_1) > 0.46$ and $(\zeta_0 + \zeta_1) > 0.71$, respectively. Thus, the rate of entropy production is negative for large values of $(\zeta_0 + \zeta_1)$ in early epoch and positive in the later epoch. Hence, the entropy law violates in the early epoch and obeys in the later epoch for these values of $(\zeta_0 + \zeta_1)$. We have also studied the generalized second law of thermodynamics. It has been observed that it is always valid for $\zeta_0 > 0$ and $0.27 \leq \zeta_1 \leq 0.85$ when λ is positive, whereas it holds for $\zeta_0 > 0$ and $0.39 \leq \zeta_1 \leq 1$ when λ is negative which are shown in Figs. 13 and 14, respectively. We have also noticed that in both the cases, the total entropy corresponds to zero in the late time of the evolution.

4. Conclusion

In this paper, we have explored the bulk viscous new HDE model in modified $f(R, T)$ gravity with bulk viscosity of the form $\zeta = \zeta_0 + \zeta_1 H$. We have analyzed the different possible phases of the Universe according to the

dimensionless bulk viscous parameters ζ_0 and ζ_1. We have calculated the scale factor, DP, statefinder and Om diagnostics, effective EoS parameter, entropy and generalized second law of thermodynamics to observe the evolution of the Universe.

In cosmology, the idea of viscous dark energy models has been presented in different ways to understand the evolution of the Universe. The notion of bulk viscosity has also been extensively studied in modified theories. This paper has explored the behavior of viscosity by considering dust matter and new HDE in the background of modified $f(R, T)$ gravity. The above investigations show that this model predicts an early deceleration followed by late time acceleration. We can conclude that the dark energy era may be explained by assuming the matter content as a non-perfect viscous fluid.

Acknowledgements We acknowledge our sincere thanks to the referees for their kind comments/suggestions to improve the manuscript. One of the authors MS would like to thank University Grant Commission, India, for providing Senior Research Fellowship (SRF) under UGC-NET scholarship.

References

[1] A G Riess et al [Supernova Search Cosmology Project Collaboration], *Astrophys. J.* **116** 1009 (1998)
[2] S Permutter et al [Supernova Cosmology Project Collaboration], *Astrophys. J.* **517** 565 (1999)
[3] D N Spergel et al *Atrophys. J. Suppl. Ser.* **148** 175 (2003)
[4] D N Spergel et al *Atrophys. J. Suppl. Ser.* **170** 377 (2007)
[5] M Tegmark et al *Phys. Rev. D* **69** 103501 (2004)
[6] G t'Hooft *Conf. Proc. C* **930308** 284 (1993); [gr-qc/9310026]
[7] R Bousso *Rev. Mod. Phys.* **74** 825 (2002)
[8] A G Cohen, D B Kaplan and A E Nelson *Phys. Rev. Lett.* **82** 4971 (1999)
[9] M Li *Phys. Lett. B* **603** 1 (2004)
[10] S D H Hsu *Phys. Lett. B* **594** 13 (2004)
[11] L Xu, W Li and J Lu *Eur. Phys. J. C* **60** 135 (2009)
[12] M Li, X D Li, S Wang and X Zhang *JCAP* **06** 036 (2009)
[13] C Gao, F Wu and X Chen *Phys. Rev. D* **79** 043511 (2009)
[14] L N Granda, A Oliveros *Phys. Lett. B* **669** 275 (2008)
[15] Y Wang and L Xu *Phys. Rev. D* **81** 083523 (2010)
[16] L N Granda and A Oliveros *Phys. Lett. B* **671** 199 (2009)
[17] K Karami and J Fehri *Int. J. Theor. Phys.* **49** 1118 (2010)
[18] F Yu, J Zhang, J Lu, W Wang and Y Gui *Phys. Lett. B* **688** 263 (2010)
[19] M Malekjani, A Khodam-Mohammadi and N Nazari-pooya *Astrophys. Space Sci.* **332** 515 (2011)
[20] J D Barrow *Phys. Lett. B* **180** 335 (1986)
[21] I Brevik and S D Odintsov *Phys. Rev. D* **65** 067302 (2002)
[22] I Brevik and O Gorbunova *Gen. Relativ. Gravit.* **37** 2039 (2005)
[23] D J Liu and X Z Li *Phys. Lett. B* **611** 8 (2005)
[24] C Eckart *Phys. Rev.* **58** 919 (1940)
[25] L D Landau and E M Lifshitz *Fluid Mechanics* (Butterworth Heinemann) (1987)
[26] W Israel and J M Stewart *Phys. Lett.* **58 A** 213 (1976)
[27] S Nojiri and S D Odintsov *Phys. Rev. D* **72** 023003 (2005)
[28] S Capozziello, V F Cardone, E Elizalde, S Nojiri and S D Odintsov *Phys. Rev. D* **73** 043512 (2006)
[29] J Ren and X H Meng *Phys. Lett. B* **633** 1 (2006)
[30] M G Hu and X H Meng *Phys. Lett. B* **635** 186 (2006)
[31] N Cruz, S Lepe and F Pena *Phys. Lett. B* **646** 177 (2007)
[32] X H Meng, J Ren and M G Hu *Commun. Theor. Phys.* **47** 379 (2007)
[33] C P Singh, S Kumar and A Pradhan *Class. Quantum Gravit.* **24** 455 (2007)
[34] C P Singh *Pramana J. Phys.* **71** 33 (2008)
[35] C J Feng and X Z Li *Phys. Lett. B* **680** 355 (2009)
[36] Y D Xu, Z G Huang and X H Zhai *Astrophys. Space Sci.* **337** 493 (2012)
[37] C P Singh and P Kumar *Eur. Phys. J. C* **74** 3070 (2014)
[38] P Kumar and C P Singh *Astrophys. Space Sci.* **357** 120 (2015)
[39] J S Gagnon and J Lesgourgues *J Cosmol. Astropart. Phys.* **09** 026 (2011)
[40] S Nojiri and S D Odintsov *Phys. Rep.* **505** 59 (2011)
[41] K Bamba, S Capozziello, S Nojiri and S D Odintsov *Astrophys. Space Sci.* **342** 155 (2012)
[42] T Harko, F S N Lobo, S Nojiri and S D Odintsov *Phys. Rev. D* **84** 024020 (2011)
[43] M Sharif and M Zubair *J. Cosmol. Astropart. Phys.* **03** 028 (2012)
[44] S Chakraborty *Gen. Relativ. Gravit.* **45** 2039 (2013)
[45] T Harko *Phys. Rev. D* **90** 044067 (2014)
[46] C P Singh and V Singh *Gen. Relativ. Gravit.* **46** 1696 (2014)
[47] E H Baffou, A V Kpadonou, M E Rodrigues, M J S Houndjo and J Tossa *Astrophys. Space Sci.* **356** 173 (2015)
[48] F Adabi, K Karami, F Felegary and Z Azarmi *Res. Astron. Astrophys.* **12** 26 (2012)
[49] K. Karami and K. Fahimi *Class. Quantum Grav.* **30** 065018 (2013)
[50] M Srivastava and C P Singh *Astrophys. Space Sci.* **363** 117 (2018)
[51] X H Meng and X Dou *Commun. Theor. Phys.* **52** 377 (2009)
[52] A Avelino, U Nucamendi *J. Cos. Astropart. Phys.* **08** 009 (2010)
[53] V Sahni, T D Saini, A A Starobinsky and U Alam *JETP Lett.* **77** 201 (2003)
[54] U Alam, V Sahni, T D Saini and A A Starobinsky *Mon. Not. R. Astron. Soc.* **344** 1057 (2003)
[55] Y B Wu, S Li, M H Fu and J He *Gen. Relativ. Gravity* **39** 653 (2007)
[56] D J Liu and W Z Liu *Phys. Rev. D* **77** 027301 (2008)
[57] C J Feng *Phys. Lett. B* **670** 231 (2008)
[58] V Sahni, A Shafieloo and A A Starobinsky *Phys. Rev. D* **78** 103502 (2008)
[59] P A R Ade et al. *Astron. Astrophys.* **594** A13 (2016)
[60] S Weinberg *Gravitation and Cosmology: principles and applications of the general theory of relativity*, (New York: Wiley) (1972)
[61] G W Gibbons, S W Hawking *Phy. Rev. D* **15** 2738 (1977)
[62] T K Mathew, R Aiswarya and K S Vidya *Eur. Phys. J. C* **73** 2619 (2013)
[63] K Karami, A Sheykhi, N Sahraei and S Ghaffari *Europhys. Lett.* **93** 29002 (2011)
[64] A Sasidharan, T K Mathew *Eur. Phys. J. C* **75** 348 (2015)
[65] M R Setare, A Sheykhi *Int. J. Mod. Phys. D* **19** 1205 (2010)
[66] S F Wu, B Wang, G H Yang and P M Zhang *Class. Quantum Grav.* **25** 235018 (2008)
[67] M Sharif, M Zubair *J. Exper. Theor. Phys.* **117** 248 (2013)
[68] D Momeni, P H R S Moraes and R Myrzakulov *Atro. Space Sci.* **361** 228 (2016)
[69] T K Mathew, M B Aswathy and M Manoj *Eur. Phy. J. C* **74** 3188 (2014)
[70] E Komatsu et al. *Astrophys. J. Suppl.* **192** 18 (2011)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

arXiv:1912.05406v1 [math.CO] 11 Dec 2019

On the resolution of the Sensitivity Conjecture

Rohan Karthikeyan, Siddharth Sinha, Vallabh Patil *

December 12, 2019

Contents

1 Introduction **2**

2 Notation **4**

3 Some Complexity Measures of Boolean Functions **4**

4 Chung's result **5**
4.1 Proving the Upper bound 6

5 Nisan-Szegedy's Result **9**
5.1 Introduction 9
5.2 Some Fourier Analysis 10
5.3 Proving Theorem 5.1 11
5.4 Proving Theorems 5.2 and 5.3 12

6 Gotsman-Linial's Result **15**
6.1 The Theorem 16
6.2 Concluding Remarks 18
6.2.1 Rubinstein's function 18

7 Hao's result **19**
7.1 Some Motivation 19
7.2 Cauchy's Interlace Theorem 19
7.3 Proving Hao's Theorem 20
7.4 Constructing the Matrix 21

8 Concluding Remarks **22**

References **22**

*Department of Applied Mathematics, Delhi Technological University, Delhi, India

2010 *Mathematics Subject Classification.* Primary 68R10, 05C35, 05-02; Secondary 68Q17, 94C10, 41A10, 42A16, 05E05, 15A24

Abstract

The Sensitivity Conjecture is a long-standing problem in theoretical computer science that seeks to fit the sensitivity of a Boolean function into a unified framework formed by the other complexity measures of Boolean functions, such as block sensitivity and certificate complexity. After more than thirty years of attacks on this Conjecture, it was finally proved in early July 2019 by Prof. Hao Huang. In this survey, we explore the ideas that inspired the proof of this Conjecture by an exposition of four papers that had the most impact on the Conjecture (including Hao's result).

1 Introduction

There are several useful measures to describe the complexity of a Boolean function that can be stated mathematically. Two such measures are sensitivity and block sensitivity. The *sensitivity* $s(f)$ of a Boolean function $f : (\mathbf{Z}/2\mathbf{Z})^n \to \{-1, 1\}$ on n variables is defined to be the largest number for which there is an x such that there are at least $s(f)$ values of $i = 1, \cdots, n$ with $f(x + e_i) \neq f(x)$. Here e_i is the Boolean vector with exactly one 1 in the i-th position. Further, the *block sensitivity*, $bs(f)$ is defined to be the maximum number of disjoint subsets of $B_1, \cdots, B_t$ of $[n] = \{1, \cdots, n\}$ such that for all $1 \leq j \leq t, f(x) \neq f(x^{B_j})$ where x^{B_j} is the Boolean string which differs from x on exactly the bits of B_j.

Now, the latter measure is known to be polynomially related to a number of other complexity measures of f, including the decision-tree complexity, and the certificate complexity. A long-standing open question that existed was whether sensitivity also belonged to this equivalence class. A positive answer to this question is commonly known as the *Sensitivity Conjecture* first proposed by Nisan and Szegedy [3].

It may be useful to think of sensitivity of a Boolean function in the following way: Suppose there is an array of n switches, with some wiring for a single light bulb. For different configurations of switches flipped on or off, the light bulb is either on, or off. Then the circuit is said to be sensitive, or *critical*, with respect to the i-th switch if for some configuration of the states of the switches, flipping the i-th switch will change the state of the light bulb from on to off, or vice versa. For that configuration, there may be more than one such switch for which the circuit is sensitive. If we count the number of such switches at which the circuit is sensitive, and do so for every configuration of the switches, then the greatest number of switches for which the circuit was sensitive for some configuration, is its global sensitivity, or its *critical complexity*.

The long-standing question for the last thirty years has been whether sensitivity is polynomially related to n, the number of input variables to the Boolean function. To resolve this, Gotsman and Linial [1] proved the equivalence between the following two problems:
1. Denote the n-dimensional cube by $Q_n = \{-1, 1\}^n$ and the maximal degree of any graph G by $\Delta(G)$. Now, for an induced subgraph G of Q_n with strictly greater than half the number of vertices in Q_n, i.e., greater than 2^{n-1} vertices, can $\Delta(G)$ be bounded below by a function of n?

2. Let $f : Q_n \to \{+1, -1\}$ be a Boolean function, with its sensitivity defined above. Denoting the degree of the multilinear polynomial representation of $f(x)$ by $\deg(f)$, is $\deg(f)$ bounded above by a function of $s(f)$?

The equivalence established by Gotsman and Linial was predicated of the representation of a Boolean function by the unique 2-colorings of the vertices of the n-dimensional hypercube. The two possible colors correspond to either -1 or 1. It was previously established by Chung et. al. [2] that for an induced sub-graph G of Q_n with strictly more than half the vertices of the Q_n, and for some vertex $v \in G$, the degree of v in G is bounded above logarithmically in n.

Recently, this bound [4] was improved from the logarithmic bound to a polynomial relation.

Theorem 1.1 (Hao)**.** *For every integer $n \geq 1$, let G be an arbitrary $(2^{n-1}+1)$-vertex induced subgraph of Q_n, then*

$$\Delta(G) \geq \sqrt{n}$$

This result directly resolves the Sensitivity conjecture:

Theorem 1.2 (Sensitivity Conjecture)**.** *There exists an absolute constant $C > 0$, such that for every Boolean function f,*

$$bs(f) \leq s(f)^C$$

Using the corollary:

Theorem 1.3. *For every Boolean function f,*

$$s(f) \geq \sqrt{\deg(f)}$$

It is found that $bs(f) \leq s(f)^4$.

In the past three decades, a great amount of research has been done on resolving the Sensitivity Conjecture, resulting in a wide-ranging body of knowledge. In this survey, we could not hope to cover even a fraction of it. The selection of material is largely limited to results that have had the most impact on the development and on the final proof of the Conjecture. We would like to direct the attention of the interested reader to the surveys of Buhrman and de Wolf [9], and Hatami, Kulkarni and Pankratov [10] for more in-depth discussions on the Conjecture.

In this survey, we will expound on the progression of results by Chung, Gotsman and Linial, Nisan, and ultimately, Hao, and the connections between their results. In the first two sections, we first develop some notation and delve into the complexity measures of Boolean functions. In section 4, the combinatorial proof of the logarithmic bound established by Chung et al., will be analysed with special consideration to its connections with the geometry of the hypercube. We will also be looking at Nisan's results characterising the degree of Boolean functions in terms of combinatorial properties of those functions in section 5. Furthermore, in section 6, we will discuss Gotsman and Linial's proof to the equivalence between the problem of the order of hypercubes and the degrees of their induced subgraphs, and the relation

between the degree and sensitivity of Boolean functions. Lastly, in section 7, we will discuss Hao's succinct proof and conclude with a discussion of the relvance of the Sensitivity Conjecture to CREW PRAMs.

We would like to thank our project advisor, Dr. Chandra Prakash Singh for his constant encouragement in the preparation of this manuscript. We would also like to thank the Quanta magazine for introducing us to this wonderful topic of sensitivity. Last but not the least, we would like to thank our parents for providing constant moral support.

2 Notation

Consider the set $Q_n = \{-1, 1\}^n$. The hypercube of dimension n is the graph with vertex set Q_n and an edge between $x = (x_1, x_2, \cdots, x_n)$ and $y = (y_1, y_2, \cdots, y_n)$ in Q_n if there is exactly one i such that $x_i \neq y_i$.

Consider a graph $G = (V, E)$. For $S \subset V$, we define an *induced subgraph* $G[S]$ to be the graph with vertex set S and edge set E' where E' consists of all edges in E that have both endpoints in S.
We denote the *maximum degree* $\Delta(G)$ of a graph $G = (V, E)$ as:

$$\Delta(G) = \max_{v \in V(G)} \deg_G(v)$$

Let's also denote $\Gamma(G) = \max(\Delta(G), \Delta(Q_n \setminus G))$. We denote by $\lambda_1(G)$ the largest eigenvalue of the graph G's adjacency matrix. And, given a real $n \times n$ matrix A, a *principal submatrix* of A is obtained by deleting the same set of rows and columns of A.

3 Some Complexity Measures of Boolean Functions

Let f be a function mapping Q_n to $\{-1, 1\}$. The *sensitivity* of f at x, denoted $s(f, x)$ is defined as the number of neighbours y of x for which $f(x) \neq f(y)$. Or mathematically stating, it's the number of i such that:

$$f(x_1, \cdots, x_i, \cdots, x_n) \neq f(x_1, \cdots, -x_i, \cdots, x_n)$$

The *(overall) sensitivity* of f, also called the critical complexity, denoted $s(f)$, is the maximum over all x in Q_n of the sensitivity of f on x, that is,

$$s(f) = \max_{x \in Q_n} s(f, x)$$

We now move on to the second complexity measure. For a string $x \in \{0, 1\}^n$ and a set $S \subseteq [n] := \{1, 2, \cdots, n\}$, we define $x^{(S)}$ to be the Boolean string which differs from x on exactly the bits in S. For a Boolean function f, we define the *block sensitivity* of f, denoted by $bs(f)$, as the maximum number t such that there exists an input $x \in \{0, 1\}^n$ and t disjoint subsets $S_1, \cdots, S_t \subset [n]$ such that for all $1 \leq i \leq t$, $f(x) \neq f(x^{(S_i)})$.

Before proceeding on further, it will be worthwhile to have a look into the decision tree model. In this model, we would like to compute the value of a given function at an unknown input. To do so, we collect information about the input by asking questions. Considering Boolean functions, we will ask only binary queries, that is, the possible answers to the query will be 0 and 1. Now, each question asked depends only on the information gained so far.
A decision tree (deterministic) can then be defined as a rooted binary tree with labels on each node and edge. Each inner node is labelled by a query. One of the two edges leaving the node is labeled 0, the other is labelled 1. The two labels represent the two possible answers to the query. Each leaf, labelled 0 or 1, give the output.

The Boolean function computed by the given decision tree takes the label at this leaf as the value on the given input. Now, we define $\text{cost}(A, x)$ as the number of queries asked when the decision tree A is executed on input x. That is, it's the length of the computation path forced by x. Also, $\max_x \text{cost}(A, x)$ defines the worst case complexity of A, that is, the tree's depth.

Now, we can define the *decision tree complexity* of the Boolean function f, denoted by $D(f)$ is given by:

$$D(f) = \min_A \max_x \text{cost}(A, x)$$

It is obvious that any function $f : \{0,1\}^n \to \{0,1\}$ can be computed by asking n questions.

And, moving towards the last complexity measure, define a b-certificate (for $b \in \{0,1\}$) as a partial assignment $p : S \to \{0,1\}$, which forces the value of f to b. Here, S is some subset of the concerned variables. The *certificate complexity* of a Boolean function f on x, denoted $C_x(f)$, is the size of the smallest certificate that agrees with x. Now, the certificate complexity of f, is defined as $C(f) = \max_{x \in Q_n} C_x(f)$. How can we visualise this measure? We can think of a Boolean function as coloring the vertex on a n-dimensional hypercube using just two colors. Then, the certificate complexity is n minus the dimension of the largest monochromatic hypercube in Q_n.

4 Chung's result

Aim: In this section, we expound Chung's result that states: An induced subgraph of the n-dimensional cube graph Q_n with strictly more than half of its vertices has a maximum degree of at least $o(\log n)$. Also, they constructed a $(2^{n-1}+1)$-vertex induced subgraph whose maximum degree is $\lceil \sqrt{n} \rceil$.
The theorem we want to prove is:

Theorem 4.1. *(Chung et.al) Let G be an induced subgraph of Q_n with at least $2^{n-1}+1$ vertices. Then for some vertex v of G, we have:*

$$\deg_G(v) > \frac{1}{2}\log n - \frac{1}{2}\log\log n + \frac{1}{2} \tag{1}$$

Also, there exists a $(2^{n-1}+1)$ vertex induced subgraph G of Q_n with

$$\Delta(G) < \sqrt{n} + 1 \tag{2}$$

4.1 Proving the Upper bound

Let's first denote $\{1, 2, \cdots, n\} := [n]$. To prove the upper bound, we consider $[n]$ and a family of sets $2^{[n]}$, that is, the set of all subsets of $[n]$. We observe that the natural bijection between the set of n-bit vectors $\{0,1\}^n$ and $2^{[n]}$ is the map $\psi : \{0,1\}^n \leftrightarrow 2^{[n]}$ such that the vector $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ is mapped to $S = \{i \mid x_i = 1, 1 \leq i \leq n\}$ and vice versa. In other words, the ith bit of $\mathbf{x}$ denotes the inclusion of the ith natural number in the subset of $[n]$; if $x_i = 1$, then S includes i.

Therefore, we may extend the bijection between the n-bit vectors and $2^{[n]}$ to the construction of Q_n. We partition Q_n into two vertex-disjoint subgraphs G and G' where G has $2^{n-1} + 1$ vertices. We consider now a certain family of sets $\mathbf{X}$ such that $|\mathbf{X}| = 2^{n-1} \pm 1$, where $\mathbf{X} \subset 2^{[n]}$. We define the family of sets $\mathbf{X}$ over a finite collection of subsets $\mathbf{F} \subset 2^{[n]}$ as:

$$\begin{aligned}\mathbf{X}(\mathbf{F}) =& \{S \subset [n] \,:\, |S| = even, \exists\, F \in \mathbf{F} \text{ with } F \subset S\} \\ & \cup \{S \subset [n] \,:\, |S| = odd, F \setminus S \neq \phi \text{ for all } F \in \mathbf{F}\}\end{aligned}$$

That is, $\mathbf{X}(\mathbf{F})$ is the collection of all even sets which contain some $F \in \mathbf{F}$ along with all odd sets which contain no $F \in \mathbf{F}$.

Let us consider $\mathbf{F}$ as the partition of $[n]$ given by $[n] = F_1 \cup F_2 \cdots \cup F_k$ such that $|k - \sqrt{n}| < 1$ and $||F_i| - \sqrt{n}| < 1$, $1 \leq i \leq k$. We now prove the existence of such a partition:

Proof. Let $k = \lceil \sqrt{n} \rceil$, and this implies that $|k - \sqrt{n}| < 1$. Now, denote $u := \sqrt{n} \implies u^2 = n$. It's easy to see that there exists an ϵ with $0 \leq \epsilon < 1$ so that $\lfloor \sqrt{n} \rfloor = \sqrt{n} - \epsilon = u - \epsilon$ and $\lceil \sqrt{n} \rceil = \sqrt{n} - \epsilon + 1 = u - \epsilon + 1$.

For the above partition to exist, we simply need to show that $\exists\, x, y \in \mathbf{Z}^+$ such that $x\lfloor \sqrt{n} \rfloor + y\lceil \sqrt{n} \rceil = n$ with $x + y = \lceil \sqrt{n} \rceil = u - \epsilon + 1$. Now, note that

$$x(u - \epsilon) + y(u - \epsilon + 1) = u^2 \tag{3}$$

must have integer solutions.

Now, observe that:

$$\begin{aligned}(3) &\implies x(x + y - 1) + y(x + y) = u^2 \\ &\implies x^2 + y^2 + 2xy - x = u^2 \\ &\implies (x + y)^2 - x = u^2 \\ &\implies (x + y)^2 = u^2 + x\end{aligned}$$

Clearly, this has integer solutions as we may fix u^2 and x and so vary y to get a set of solutions. □

For such a partition of $[n]$, we wish to see if the following claims are true:

1. $|\mathbf{X}| = 2^{n-1} \pm 1$.

2. For the subgraph induced by $\mathbf{X}(\mathbf{F})$ and $2^{[n]} \setminus \mathbf{X}(\mathbf{F})$,

$$\Delta \leq k$$

Let the reader allow us to introduce some terminology here. We define the *rank* of $\mathbf{F}$ as the largest size of an element in $\mathbf{F}$ and denote it $r(\mathbf{F})$. Also, we denote $t(\mathbf{F})$ as the largest size of the disjointly representable subsystems of $\mathbf{F}$. That is, $t(\mathbf{F})$ is the maximum value of t such that one can find $F_1, F_2, \cdots, F_t \in \mathbf{F}$ and $x_i \in F_i$, $1 \leq i \leq t$, such that $x_i \in F_j \leftrightarrow i = j$, for all $1 \leq i, j \leq k$, that is, no F_i is contained in the union of the others.

In the partition chosen for $\mathbf{F}$ such that we had k disjointly representable subsets of $\mathbf{F}$, we note that $t(\mathbf{F}) = k$ and $r(\mathbf{F})$ is either $\lfloor \sqrt{n} \rfloor$ or $\lceil \sqrt{n} \rceil$ (How?). Since, the degree of a vertex corresponding to some $S \in \mathbf{X}(\mathbf{F})$ is the number of $S_i \in \mathbf{X}(\mathbf{F})$ such that S and S_i differ by only one element, the maximum degree $\Delta(G(\mathbf{F}))$ is either bounded by $r(\mathbf{F})$ or $t(\mathbf{F})$. We can then easily see that $\Delta(G(\mathbf{F})) < \sqrt{n}$.

Generally, we can see that:

Lemma 4.2. $\Delta(G(\boldsymbol{F})) \leq \max\{r(\boldsymbol{F}), t(\boldsymbol{F})\}$

Remark 1. The same result also holds true for $\Delta(G'(\mathbf{F}))$.

Proof. An interesting thing to note is that this result holds for any family of subsets of $2^{[n]}$. To see this, consider the sets corresponding to the vertices: (1 1 0 1 1), (1 1 0 1 0), (0 1 0 1 1) of $G(\mathbf{F})$ for some $\mathbf{F} \subset 2^{[n]}$. The subsets $\{1, 2, 4\}$ and $\{2, 4, 5\}$ belong to $\mathbf{X}(\mathbf{F})$ so long as no subset of either belongs to $\mathbf{F}$. Therefore, there are no even subsets in $\mathbf{F}$ which are contained in odd sets of $\mathbf{X}(\mathbf{F})$.

Consequently, (1 1 1 1 1) cannot belong to $G(\mathbf{F})$ as it would contain some $F \in \mathbf{F}$. Consider again the vertex (1 1 0 1 1). It can have maximum degree 4 (in the 5-cube graph) if (0 1 0 1 1), (1 0 0 1 1), (1 1 0 0 1) and (1 1 0 1 0) belong to $G(\mathbf{F})$. However, if no subsets of the vertices adjacent to (1 1 0 1 1) belong to $\mathbf{F}$, then only the singleton elements (1 0 0 0 0), (0 1 0 0 0), $\cdots$, (0 0 0 0 1) may belong to $G(\mathbf{F})$, each with degree at most 1.

Therefore, for any edge of Q_n corresponding to (S, S'), where $S, S' \in \mathbf{X}(\mathbf{F})$ and S is even, then the odd set S' can be seen to be a proper subset of S. Since, S and S' differ by the inclusion of precisely one element of $[n]$, and the exclusion of that element from S results in a subset not contained in $\mathbf{F}$, then

$$\deg(S) \leq \left| \bigcap \{F : F \in \mathbf{F}, F \subset S\} \right| \leq r(\mathbf{F}) \tag{4}$$

Taking S as the odd set, we instead get $S \subset S'$, and by the construction of $G(\mathbf{F})$, there exists a $F \in \mathbf{F}$ such that $F \subset S'$ and $F \not\subset S$. Suppose, we have the sets $S'_1, S'_2, \cdots, S'_a$ such that there is an edge between S and S'_i, $1 \leq i \leq a$. Then, there must also be some sets $F_1, F_2, \cdots, F_a \in \mathbf{F}$ which don't belong to S, but $F_i \in S'_i$. Since, S and S'_i differ by the inclusion of exactly one element $F_1, F_2, \cdots, F_a$ must be disjointly representable. Hence, $a \leq t(\mathbf{F})$. Since the number a gives the number of adjacent vertices, we have completed the proof of the lemma. □

Now, the cardinality of $\mathbf{X}(\mathbf{F})$ can be found using the sieve formula and is given as:

$$|\mathbf{X}(\mathbf{F})| = 2^{n-1} + (1-2\epsilon)\Bigg[\sum_{\substack{F_i \in \mathbf{F} \\ |F_i|=n}} 1 - \sum_{\substack{F_i \in \mathbf{F} \\ |F_i \cup F_j|=n}} \sum_{F_j \in \mathbf{F}} 1 + \cdots\Bigg] \tag{5}$$

where $\epsilon = 0$ when n is even, and equals 1 when odd.

Denote by $f(\mathbf{F})$ the bracketed expression on the RHS of (5). When we take $\mathbf{F}$ as the partition of $[n]$ into $F_1, F_2, \cdots, F_k$, $f(\mathbf{F})$ can be easily seen to equal $(-1)^{k+1}$. And, also we can easily verify that $(1-2\epsilon) = (-1)^n$ to give:

$$|\mathbf{X}(\mathbf{F})| = 2^{n-1} + (-1)^{n+k+1}$$

thus proving claim 1.

Now, to prove claim 2, since we are concerned only with partitions where one part contains more than half of the vertices, we consider the case where $f(\mathbf{F}) \neq 0$, and prove this lemma:

Lemma 4.3. *Suppose $f(\boldsymbol{F}) \neq 0$. Then, $\max\{r(\boldsymbol{F}), t(\boldsymbol{F})\} \geq \sqrt{n}$.*

Remark 2. We choose the maximum of $r(\mathbf{F})$ and $t(\mathbf{F})$ since the degree is bounded by both values, and the greater value is the one which dominates the bound.

Proof. Observe that the motivation for $f(\mathbf{F}) \neq 0$ stems from our choice of a partition in which one subgraph of the hypercube contains strictly greater than half the vertices. Taking strictly more than half of the vertices implies that there is a vertex in every direction. Take any n-bit vector. Now, there are 2^{n-1} possible n-bit vectors such that every one of them has their jth component as 0. Since, we are taking $2^{n-1}+1$ vertices, by the pigeon-hole argument, there is at least one vector in every direction, that is, has a non-zero value for the jth component. Since vertices of $G(\mathbf{F})$ correspond to the subsets of $2^{[n]}$, this implies that $\left|\bigcap \mathbf{F}\right| = \sqrt{n}$.

Let us choose a subfamily $\{F_1, F_2, \cdots, F_s\}$ of $\mathbf{F}$ such that this family has the smallest number of sets such that $\bigcup_{i=1}^{s} F_i = [n]$. Now, $r(\mathbf{F}) < \sqrt{n}$, else the inequality $\max\{r(\mathbf{F}), t(\mathbf{F})\} \geq \sqrt{n}$ is trivial for $|F| \geq \sqrt{n}$. Then, we must check if $t(\mathbf{F}) \geq \sqrt{n}$ in such a case. Since, s is the minimum number such that $\bigcup_{i=1}^{s} F_i = [n]$, so $\{F_1, F_2, \cdots, F_s\}$ must be disjointly representable, else there would be a contradiction to the minimality of S.

Since $|F_i|$ cannot exceed $\sqrt{n}$, we clearly see that if $\max|F_i| < \sqrt{n}$, and if $s < \sqrt{n}$, then: $s \cdot \max\{|F_i|\} < |\,[n]\,| = n$, which is a contradiction since $\bigcup_{i=1}^{s} F_i = [n]$. Therefore $s \geq \sqrt{n}$, and so consequently:

$$\sqrt{n} \leq \max\{r(\mathbf{F}), t(\mathbf{F})\}$$

Since $\sqrt{n} < n$ for $n \geq 1$, the maximal degree $\Delta(G) \leq \sqrt{n}$, and since both $\Delta(G)$ and $\sqrt{n}$ are bounded above by $\max\{r(\mathbf{F}), t(\mathbf{F})\}$, we can have a partition $G(\mathbf{F})$ such that $\Delta(G) \ll \sqrt{n}$. □

5 Nisan-Szegedy's Result

Aim: In this section, we will look at Nisan, Szegedy's results that establish a polynomial relation between the decision tree complexity of Boolean functions, the degree of the multilinear polynomial representing it, and the smallest degree of the polynomial function approximating the Boolean function.

5.1 Introduction

The Sensitivity conjecture was formally proposed by Nisan and Szegedy after characterising the degree of Boolean functions represented by multilinear functions in terms of the combinatorial properties of the Boolean functions. This is done by representing the Boolean function as polynomials over the real numbers, $\mathbf{R}$. If we define a Boolean function f as $f : \{F, T\}^n \to \{F, T\}$, then we can encode T as 1 and F as 0, thus creating a mapping from a subset of $\mathbf{R}^n$ to a subset of $\mathbf{R}$. Let us denote such a polynomial as p.

A real multivariate polynomial $p : \mathbf{R}^n \to \mathbf{R}$ represents f if for all $x \in \{0, 1\}^n$, $f(x) = p(x)$. There is always a unique multilinear polynomial representing a given Boolean function. The results established by Nisan and Szegedy concern themselves with the most basic parameter of representing Boolean functions as polynomials, its degree.
We define the *degree* of a Boolean function f, denoted by $\deg(f)$, as the degree of the unique multilinear real polynomial that represents f exactly.

The first result we would like to show is establishing the lower bound on the degree of a Boolean function f in terms of the number of variables n.

Theorem 5.1. *Let f be a Boolean function that depends on n variables. Then,*

$$\deg(f) \geq \log_2 n - O(\log \log n) \tag{6}$$

We next show a relation between the degree of the Boolean function f and the decision tree complexity of f along with the smallest degree of a polynomial that approximates f, denoted by $\widetilde{\deg}(f)$. Now consider a real polynomial p. We say that p *approximates* f if for every $x \in \{0, 1\}^n$, we have that

$$|p(x) - f(x)| < \frac{1}{3} \tag{7}$$

The approximate degree of f, that is, $\widetilde{\deg}(f)$ is defined to be the minimum over all polynomials p that approximate f of the degree p.
Hence, we have the following inequalities:

Theorem 5.2. *For every Boolean function f,*

$$\deg(f) \leq D(f) \leq 16 \deg(f)^8 \tag{8}$$

Theorem 5.3. *There exists a constant c such that for every Boolean function, we have:*

$$\widetilde{\deg}(f) \leq \deg(f) \leq D(f) \leq c \cdot \widetilde{\deg}(f)^8 \tag{9}$$

5.2 Some Fourier Analysis

In this subsection, we state and prove some results needed for our discussions from [5]. Consider the Fourier transform representation, that is, considering the Boolean function f as $f : \{-1,1\}^n \to \{-1,1\}$. Interpret the domain $\{-1,1\}^n$ of f as 2^n points lying in $\mathbf{R}^n$ and think of f as giving a ± 1 labeling to each of these points. There is a familiar method for interpolating such data points with a polynomial. We'll let the reader refer to [5] for more details.

Generally, after writing out the polynomial, we observe that this polynomial is always "multilinear" - that is, it has no variables squared, cubed, etc. In general, a multilinear polynomial over variables $x_1, \cdots, x_n$ has 2^n terms, one for each monomial $\prod_{i \in S} x_i$, where $S \subseteq [n] = \{1, 2, \cdots, n\}$ (Note: $\prod_{i \in \phi} x_i$ denotes 1). Hence:

Lemma 5.4. *Every function $f : \{-1,1\}^n \to \mathbf{R}$ can be expressed uniquely as a multilinear polynomial,*

$$f(x) = \sum_{S \subseteq [n]} c_S \prod_{i \in S} x_i \tag{10}$$

where each c_S is a real number.

This expression (10) is precisely the "Fourier expansion" of f. It's a convention to write the coefficient c_S as $\hat{f}(S)$ and the monomial $\prod_{i \in S} x_i$ as $\chi_S(x)$. Thus we finally have:

$$f(x) = \sum_{S \subseteq [n]} \hat{f}(S) \chi_S(x) \tag{11}$$

Let us denote $\mathbf{x} = (\mathbf{x_1}, \cdots, \mathbf{x_n})$ to denote a uniformly random string from $\{-1,1\}^n$ where each $\mathbf{x_i}$ is a random variable. We can think of generating such an $\mathbf{x}$ by choosing each bit $\mathbf{x_i}$ independently and uniformly from $\{-1,1\}$.

The result we'll need for our discussions is:

Theorem 5.5. *(Parseval) For any $f : \{-1,1\}^n \to \mathbf{R}$,*

$$\sum_{S \subseteq [n]} \hat{f}(S)^2 = \mathbf{E}_{\mathbf{x}}[f(\mathbf{x})^2]$$

Proof. By the Fourier expansion of f,

$$\begin{aligned}
\mathbf{E}_{\mathbf{x}}[f(\mathbf{x})^2] &= \mathbf{E}_{\mathbf{x}}\left[\left(\sum_{S \subseteq [n]} \hat{f}(S)\chi_S(\mathbf{x})\right)^2\right] \\
&= \mathbf{E}_{\mathbf{x}}\left[\sum_{S,T \subseteq [n]} \hat{f}(S)\hat{f}(T)\chi_S(\mathbf{x})\chi_T(\mathbf{x})\right] \\
&= \sum_{S,T \subseteq [n]} \hat{f}(S)\hat{f}(T)\mathbf{E}_{\mathbf{x}}[\chi_S(\mathbf{x})\chi_T(\mathbf{x})]
\end{aligned}$$

Recalling that $\chi_S(x)$ denotes $\prod_{i \in S} x_i$, we see that $\chi_S(x)\chi_T(x) = \chi_{S \,\Delta\, T}(x)$. This is because whenever $i \in S \cap T$, we get an x_i^2, which can be replaced by 1. So, we get

$$\sum_{S,T \subseteq [n]} \hat{f}(S)\hat{f}(T)\mathbf{E}_{\mathbf{x}}[\chi_S(\mathbf{x})\chi_T(\mathbf{x})] = \sum_{S,T \subseteq [n]} \hat{f}(S)\hat{f}(T)\mathbf{E}_{\mathbf{x}}[\chi_{S \,\Delta\, T}(\mathbf{x})]$$

We now observe that $\mathbf{E_x}[\chi_U(\mathbf{x})] = 0$, unless $U = \phi$ in which case we get a 1. This holds because by independence of the random bits $\mathbf{x_1}, \mathbf{x_2}, \cdots, \mathbf{x_n}$, we have $\mathbf{E_x}[\chi_U(\mathbf{x})] = \mathbf{E_x}[\prod_{i \in x_i} \mathbf{x_i}] = \prod_{i \in U} \mathbf{E_x}[\mathbf{x_i}]$ and each $\mathbf{E}[\mathbf{x_i}] = 0$.

Finally, we deduce that

$$\sum_{S,T \subseteq [n]} \hat{f}(S)\hat{f}(T)\mathbf{E_x}[\chi_{S \Delta T}(\mathbf{x})] = \sum_{S \Delta T = \phi} \hat{f}(S)\hat{f}(T) = \sum_{S \subseteq [n]} \hat{f}(S)^2$$

as claimed. □

Using linearity of expectation, and some facts we used in the proof above, we can see that:

Lemma 5.6. *For any $f : \{-1,1\}^n \to \mathbf{R}$ and $S \subseteq [n]$,*

$$\hat{f}(S) = \mathbf{E_x}[f(\mathbf{x})\chi_S(\mathbf{x})]$$

Finally for ordinary Boolean functions $f : \{-1,1\}^n \to \{-1,1\}$ we have $f(x)^2 = 1$ for every x; hence Parseval has the following important corollary:

Corollary 5.7. *(Parseval's equality) If $f : \{-1,1\}^n \to \{-1,1\}$, then*

$$\sum_{S \subseteq [n]} \hat{f}(S)^2 = 1$$

5.3 Proving Theorem 5.1

We will make use of an important definition due to Kalai: For a Boolean function on n variables and a variable x_i, the *influence* of x_i on f, denoted by $\mathrm{Inf}_i(f)$, is defined to be:

$$\mathrm{Inf}_i(f) = \Pr[f(x) \neq f(x^{(i)})]$$

where $x^{(i)}$ denotes the string $x \in \{0,1\}^n$ with the ith bit flipped and x is chosen uniformly in $\{\text{false}, \text{true}\}^n$. In words, $\mathrm{Inf}_i(f)$ is the probability that flipping the ith coordinate flips the value of the function. Kalai also defines a lemma in [7]:

Lemma 5.8. *For any Boolean function f on n variables, using the Fourier transform representation of f, we have:*

$$\sum_{i=1}^{n} Inf_i(f) = \sum_{S} |S|\hat{f}(S)^2 \tag{12}$$

Combining this lemma with Parseval's equality, we can conclude that:

Corollary 5.9. *For any Boolean function f,*

$$\sum_{i=1}^{n} Inf_i(f) \leq \deg(f) \tag{13}$$

Now, we need the Schwartz-Zippel lemma that gives an upper bound for the number of $\{-1,1\}$ zeroes of any multilinear polynomial.

Lemma 5.10. *Let $p(x_1, \cdots, x_n)$ be a non-zero multilinear polynomial of degree d. If we choose $x_1, \cdots, x_n$ at random in $\{-1, 1\}$, then:*

$$\Pr[p(x_1, \cdots, x_n) \neq 0] \geq 2^{-d}$$

We will leave it to the reader to prove this lemma by an induction on n and writing p as a linear combination of two polynomials.

Let's now prove (6).

Proof. For each i define a function f^i on $n-1$ variables as follows:

$$f^i(x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n) = f(x_1, \cdots, -1, x_{i+1}, \cdots, x_n) - f(x_1, \cdots, 1, x_{i+1}, \cdots, x_n)$$

Now from the definition of influence we easily get that:

$$\mathrm{Inf}_i(f) = \Pr[f^i(x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n) \neq 0]$$

where $x_1, \cdots, x_{i_1}, x_{i+1}, \cdots, x_n$ are chosen at random in $\{-1, 1\}$.

Since f depends on all the variables, we have that for every i, f^i is not identically zero, and thus we can use Lemma 5.10 and conclude that for all i, $\mathrm{Inf}_i(f) \geq 2^{-d}$. On the other hand, from (13) it follows that $\sum_i \mathrm{Inf}_i(f) \leq d$. Combining these two bounds gives us the required result. □

5.4 Proving Theorems 5.2 and 5.3

Now, we will return to the representation of true $= 1$ and false $= 0$. Also, we'll need the notion of symmetric polynomials. A polynomial $f(x_1, x_2, \cdots, x_n)$ is a symmetric polynomial if

$$f(x_1, \cdots, x_n) = f(x_{\sigma(1)}, \cdots, x_{\sigma(n)})$$

for any permutation σ of $[n] = \{1, \cdots, n\}$. We will next use the method of symmetrization.

Let $p : \mathbf{R}^n \to \mathbf{R}$ be a multivariate polynomial. Then the symmetrization of p is

$$p^{\mathrm{sym}}(x_1, \cdots, x_n) = \frac{\sum_{\sigma \in S_n} p(x_{\sigma(1)}, \cdots, x_{\sigma(n)})}{n!}$$

Since, the number of permutations of $[n]$ is $n!$, we get a sum of $n!$ terms each term resolving to give 0 or 1.
For illustration, let us consider a polynomial of two variables, $p : \{0,1\}^2 \to \mathbf{R}$ given by $p(x_1, x_2) = x_1 + x_1 x_2$. By the laws of Boolean algebra we know that: $x_1 + x_1 x_2 = x_1(1 + x_2) = x_1$. Then, we can easily get:

$$p^{\mathrm{sym}}(x_1, x_2) = \frac{x_1 + x_2}{2}$$

Hence, the important point to note here is that if we are only interested in inputs $x \in \{0,1\}^n$ then p^{sym} turns out to depend only upon $x_1 + \cdots + x_n$. We can thus represent it as a univariate polynomial of $x_1 + \cdots + x_n$ (see [11]):

Lemma 5.11. *Taking $p : \mathbf{R}^n \to \mathbf{R}$ to be a multivariate polynomial, then there exists a unique univariate polynomial $\tilde{p} : \mathbf{R} \to \mathbf{R}$ such that for all $x_1, \cdots, x_n \in \{0,1\}^n$, we have:*

$$p^{sym}(x_1, \cdots, x_n) = \tilde{p}(x_1 + \cdots + x_n)$$

Moreover, $\deg(\tilde{p}) \leq \deg(p)$.

Proof. Let the degree of p^{sym} be d. Let P_k denote the sum of all $\binom{n}{k}$ products $\prod_{i \in S} x_i$ of the $|S| = k$ different variables. Since p^{sym} is symmetric, it can be shown by induction that it can be written as:

$$p^{\text{sym}}(x) = c_o + c_1 P_1(x) + c_2 P_2(x) + \cdots + c_d P_d(x)$$

with $c_i \in \mathbf{R}$. Observe that on $x \in \{0,1\}^n$ with $z := x_1 + \cdots + x_n$ ones, P_k assumes value:

$$P_k(x) = \binom{z}{k} = \frac{z(z-1)\cdots(z-k+1)}{k!}$$

which is a polynomial of degree k of z. Therefore the univariate polynomial $\tilde{p}(z)$ defined by:

$$\tilde{p}(z) : +c_0 + c_1 \binom{z}{1} + c_2 \binom{z}{2} + \cdots + c_d \binom{z}{d}$$

has the desired property. □

As $\deg(p)$ is bounded below by $\deg(\tilde{p})$, we wish to find a bound for $\deg(\tilde{p})$. We will therefore need to use a result from approximation theory (see, e.g. [12]).

Theorem 5.12. *(Markov Inequality) Let $p : \mathbf{R} \to \mathbf{R}$ be a univariate polynomial of degree d that for any real number $a_1 \leq x \leq a_2$ satisfies $b_1 \leq p(x) \leq b_2$. Then, for all $a_1 \leq x \leq a_2$, the derivative of p satisfies*

$$|p'(x)| \leq \frac{d^2(b_2 - b_1)}{a_2 - a_1}$$

In our case, because we have information on the values of $p(x)$ only for integer x, we need this lemma next:

Lemma 5.13. *Let p be a polynomial with the following properties:*

1. *For any integer $0 \leq i \leq n$, we have $b_1 \leq p(i) \leq b_2$.*
2. *For some real $0 \leq x \leq n$, the derivative of p satisfies $|p'(x)| \geq c$.*

Then

$$\deg(p) \geq \sqrt{\frac{cn}{c + b_2 - b_1}}$$

Proof. Let $c' = \max_{0 \leq x \leq n} |p'(x)|$ which is therefore definitely greater than or equal to c. It is also clear that for all real $0 \leq x \leq n$:

$$b_1 - \frac{c'}{2} \leq p(x) \leq b_2 + \frac{c'}{2}$$

Using the Markov inequality we have

$$c' \leq \frac{\deg(p)^2(c' + b_2 - b_1)}{n}$$

Thus,

$$\deg(p)^2 \geq \frac{c'n}{c' + b_2 - b_1} \geq \frac{cn}{c + b_2 - b_1}$$

□

Lemma 5.14. *Let f be a Boolean function such that $f(000\cdots0) = 0$ and for every Boolean vector $\mathbf{x}$ of Hamming weight 1, $f(\mathbf{x}) = 1$. Then:*

$$\deg(f) \geq \sqrt{\frac{n}{2}} \tag{14}$$

and

$$\widetilde{\deg}(f) \geq \sqrt{\frac{n}{6}} \tag{15}$$

Proof. We will first prove the bound for $\widetilde{\deg}(f)$. The sharper bound for $\deg(f)$ follows exactly the same lines. Let p be a univariate polynomial approximating f, and consider $\tilde{p}$ the univariate polynomial giving its symmetrization. Now, $\tilde{p}$ satisfies the following properties:

1. By Lemma 5.11, we have $\deg(\tilde{p}) \leq \deg(p)$.
2. Since for every Boolean vector $\mathbf{x}$, then $p(\mathbf{x})$ is within $\frac{1}{3}$ of a Boolean value (by equation (2)), so for every integer $0 \leq i \leq n$, $-\frac{1}{3} \leq \tilde{p}(i) \leq \frac{4}{3}$.
3. Since, $f(000\cdots0) = 0$, and by equation (2) once again, we get $\tilde{p}(0) \leq \frac{1}{3}$.
4. Since for all Boolean vectors $\mathbf{x}$ of Hamming weight 1, $f(\mathbf{x}) = 1$ and by equation (2) once again, $\tilde{p}(1) \geq \frac{2}{3}$.

By the properties (3) and (4) listed above and on using the Mean Value Theorem (MVT) for derivatives, we obtain for some real $0 \leq z \leq 1$, the derivative $\tilde{p}'(z) \geq \frac{1}{3}$. We can now apply Lemma 5.13 to obtain the lower bound for $\deg(\tilde{p})$. We can also get a similar bound for $\deg(p)$ by using $0 \leq \tilde{p}(i) \leq 1$ along with Lemma 5.13 and the MVT for derivatives. □

The authors also give an example of a function f satisfying $f(\mathbf{0}) = 0$ and $f(e_i) = 1$ where $\mathbf{0}$ is the zero vector and e_i is the Boolean vector with one 1 at the i-th position whose degree is significantly smaller than n.

Lemma 5.15. *There exists an (explicitly given) Boolean function f of n variables satisfying $f(\mathbf{0}) = 0$ and $f(e_i) = 1$ and $\deg(f) = n^\alpha$ for $\alpha = \log_3 2 = 0.631\cdots$.*

We leave it to the reader to prove the lemma by constructing a Boolean function $E_3(x, y, z)$ and building a complete ternary tree on the n variables by considering an extension of the function.

Now, in an earlier paper, Nisan [6] proved an polynomial relation between the block sensitivity and the decision tree complexity as:

$$bs(f) \leq D(f) \leq bs^4(f) \tag{16}$$

We can thus get lower bounds for the degree in terms of the block sensitivity.

Lemma 5.16. *For every Boolean function f,*

$$\deg(f) \geq \sqrt{\frac{bs(f)}{2}} \tag{17}$$

and

$$\widetilde{deg}(f) \geq \sqrt{\frac{bs(f)}{6}} \tag{18}$$

Proof. We refer to Theorem 14.11, [13] for the proof of this lemma. Let $f(x)$ be a Boolean function on n variables, and let $q : \mathbf{R}^n \to \mathbf{R}$ be the multilinear polynomial of degree d representing f. By Lemma 5.14, we know that every Boolean function f of n variables, which rejects the zero vector and accepts all n vectors with Hamming weight 1, has $\deg(f) \geq \sqrt{\frac{n}{2}}$. It's therefore enough to construct a multilinear polynomial p of $t = bs(f)$ variables satisfying the conditions of this lemma.

Let $t = bs(f)$, and a and $S_1, \cdots, S_t$ be the input and the sets achieving the block sensitivity. Let us assume WLOG that $f(a) = 0$. We transform $q(x_1, x_2, \cdots, x_n)$ into a multilinear polynomial $p(y_1, y_2, \cdots, y_t)$ of t new variables by replacing every variable x_j in p as follows: Define a function $f'(y_1, \cdots, y_t)$ as follows:

$$x_j := \begin{cases} y_i, & \text{if } a_j = 0 \text{ and } j \in S_i \\ 1 - y_i, & \text{if } a_j = 1 \text{ and } j \in S_i \\ a_j & \text{if } j \notin S_1 \cup \cdots \cup S_t \end{cases}$$

That is, for $y \in \{0,1\}^t$ we have that:

$$p(y) = q(a \oplus y_1 S_1 \oplus \cdots \oplus y_t S_t)$$

where

$$y_i S_i = (0, \cdots, 0, \overbrace{y_1, \cdots, y_i}^{S_i}, 0, \cdots, 0)$$

We can easily verify that p is a multilinear polynomial of degree at most d, and satisfies the conditions of Lemma 5.14. We can therefore conclude that

$$d = \deg(q) \geq \deg(p) \geq \sqrt{\frac{t}{2}} = \sqrt{\frac{bs(f)}{2}}$$

The proof of the other part is analogous. □

Since $\deg(f) \leq D(f)$ is obvious (How? Note that the tests along paths to 1-leaves define a multilinear polynomial), by combining (16) and the results of Lemma 5.16, we have (8) and (9).

6 Gotsman-Linial's Result

Aim: In this section we look at Gotsman, Linial's results that reduce the Sensitivity Conjecture to answering a simple question about cubes of different dimensions: If you choose any collection of more than half the corners of a cube and color them red, is there always some red point that is connected to many other red points?

6.1 The Theorem

Let us consider a Boolean function f. Let g be the same function as f except that we flip the value on all odd vertices. Notice now that the sensitivity of f on x is the number of i such that

$$g(x_1, \cdots, x_i, \cdots, x_n) = g(x_1, \cdots, -x_i, \cdots, x_n)$$

This can be easily visualized as follows: Take a Boolean function f and for any $x \in Q_n$, calculate $f(x)$, marking the vertex blue if $f(x) = 1$ and red, if otherwise. Thus, the sensitivity of f on x using f is defined to be the number of neighbours of x in Q_n that have a different color than x. We can also easily define sensitivity using g.

Let now G be the induced subgraph of vertices of x such that $g(x) = -1$ if the node x was colored blue and H be the induced subgraph on the set of x such that $g(x) = 1$ if the node x was colored red. The sensitivity of f, using this new notation, is defined to be the maximum number of neighbours of any vertex in G or H. Now, consider f as a multilinear polynomial over the reals which is possible when we consider f as a function going from $\mathbf{R}^n$ to $\mathbf{R}$. Now, the Sensitivity conjecture states that there is some $\alpha > 0$ such that if f has degree n, then f has sensitivity at least n^α.

Now, we note that $g(x_1, \cdots, x_n) = f(x_1, \cdots, x_n)x_1 \cdots x_n$. If f has a degree n term, the variables in that term cancel out on Q_n (because $x_i^2 = 1$) and the constant of the degree n term of f becomes the constant term of g. The constant term is just the expected value, so f has full degree iff g is unbalanced.
Now, we've seen earlier how Nisan and Szegedy show that $d(f)$ is equivalent to the other known complexity measures of Boolean functions such as: block sensitivity, certificate complexity and decision tree depth. Szegedy also independently proved a relation between $d(f)$ and $s(f)$ as

$$d(f) \geq \sqrt{s(f)}$$

Gotsman and Linial proved the following remarkable equivalence that helps in proposing an upper bound for $d(f)$ in terms of $s(f)$:

Theorem 6.1. *The following are equivalent for any function $h : \mathbf{N} \to \mathbf{R}$:*

1. *For any induced sub-graph G of Q_n such that $|V(G)| \neq 2^{n-1}$, $\Gamma(G) \geq h(n)$.*
2. *For any Boolean function f, $h(d(f)) < s(f)$.*

Remark 3. In simple graph-theoretical terms: Suppose you have a partition of the hypercube graph Q_n into sets A and B such that $|A| \neq |B|$, and let G and H be the induced sub-graphs of A and B. Then, there is some constant $\alpha > 0$ such that there is a node of A or B with atleast n^α neighbours.

Remark 4. Hao proved that given any subset A of the vertices of a hypercube with $|A| > 2^{n-1}$, the induced subgraph has a node of degree at least $\sqrt{n}$. Since either A or B in the G-L assumption has size greater than 2^{n-1}, Hao's result proves the Sensitivity conjecture.

Proof. Let's convert statement 1 of the theorem into a statement on Boolean functions: Associate with the induced subgraph G a Boolean function g such that $g(x) = 1$ iff $x \in V(G)$. How can we do this? One way would be to use Karnaugh maps. Now observe that $\deg_G(x) = n - s(g, x)$ for $x \in V(G)$ and the same relation also holds in $Q_n \setminus G$ for $x \notin V(G)$. (How?)

Let's denote by $\mathbf{E}(g) = 2^{-n} \sum_x g(x)$ the average value of g on Q_n. Now, statements 1 and 2 of the theorem can be seen as equivalent to:

Lemma 6.2. *Theorem 6.1 can be reformulated as:*

I For any Boolean function g, $\mathbf{E}(g) \neq 0 \implies \exists x : s(g, x) \leq n - h(n)$.

II For any Boolean function f, $s(f) < h(n) \implies d(f) < n$.

Proof. Seeing $1 \to I$: Note that $\Gamma(G) = \max(\Delta(G), \Delta(Q_n \setminus G))$. And also, $\Delta(G) = \max_{x \in V(G)} \deg_G(x)$. Hence, we observe that:

$$\begin{aligned} \Gamma(G) = \max(\Delta(G), \Delta(Q_n \setminus G)) \geq h(n) &\implies \Delta(G) \geq h(n) \\ &\implies \max_{x \in V(G)} \deg_G(x) \geq h(n) \\ &\implies \exists x : \deg_G(x) \geq h(n) \\ &\implies \exists x : n - s(g, x) \geq h(n) \\ &\implies \exists x : s(g, x) \leq n - h(n) \end{aligned}$$

where the first implication is done assuming that $\Delta(G) \geq \Delta(Q_n \setminus G)$.
Seeing $I \to 1$: The requirement of $\mathbf{E}(g) \neq 0$ corresponds to $|V(G)| \neq 2^{n-1}$. Now, we can just reverse the implications of the previous result to prove the equivalence of 1 and I.

Seeing $2 \to II$: Given $d(f) < h^{-1}(s(f))$, then it's easy to see that if $s(f) < h(n)$, the result follows.
Seeing $II \to 2$: To prove the reverse implication, let f be a Boolean function of degree d. Fix a monomial of degree d of the representing polynomial of f. Without loss of generality, we may assume the monomial is $x_1 \cdots x_d$. Define $g(x_1, \cdots, x_d) := f(x_1, \cdots, x_d, 0, \cdots, 0)$. Then, g has full degree d, so it follows that $s(f) \geq s(g) \geq h(d) = h(\deg f)$, as desired. □

To see the equivalence of I and II, define

$$g(x) = f(x)p(x)$$

where $p(x) = (-1)^{\sum x_i}$ is the parity function (note, we take the range of Boolean functions to be $\{-1, +1\}$). Since the parity function is sensitive to all n variables, we observe that $\forall x \in Q_n$:

$$s(g, x) = n - s(f, x) \tag{19}$$

and also for all $S \subset [n]$:

$$\hat{g}(S) = 2^{-n} \sum_x g(x) \prod_{i \in S} x_i = 2^{-n} \sum_x f(x) \prod_{i \notin S} x_i = f([n] \setminus S) \tag{20}$$

In particular,

$$\mathbf{E}(g) = \hat{g}(\phi) = \hat{f}([n]) \tag{21}$$

where $\hat{f}(S)$ denotes the Fourier transform of f at $S \subset [n]$, that is, the highest order coefficient in the representation of f as a polynomial.

Seeing $I \to II$: Assume that $d(f) = n$, that is, $\hat{f}([n]) \neq 0$. This implies by (21) that $\mathbf{E}(g) \neq 0$ and by I,

$$\exists x : s(g, x) \leq n - h(n) \implies \exists x : s(f, x) \geq h(n)$$

a contradiction.
Seeing $II \to I$: Assume

$$\forall x : s(g, x) = n - s(f, x) > n - h(n) \implies s(f) < h(n)$$

as $s(f) = \max_{x \in Q_n} s(f, x)$. Then by II,

$$d(f) < n \implies \hat{f}([n]) = \hat{g}(\phi) = \mathbf{E}(g) = 0$$

a contradiction.

□

6.2 Concluding Remarks

Now, Gotsman-Linial's result translates a Boolean function with a polynomial gap between degree and sensitivity into a graph with the same polynomial gap between Γ and n, and vice-versa. For example, observe that Rubinstein's function (given below) can be used to obtain a graph G with the surprising property $\Gamma(G) = \Theta(\sqrt{n})$. Chung et.al [2] also constructed a graph G with $\Gamma(G) < \sqrt{n} + 1$. This theorem implies that the following conjecture is equivalent to the Sensitivity Conjecture.

Lemma 6.3. *There is a constant $\epsilon > 0$ such that for every induced subgraph G of Q_n with $|V(G)| \neq 2^{n-1}$ we have $\Gamma(G) \geq n^{\epsilon}$.*

6.2.1 Rubinstein's function

The following function was constructed by Rubinstein: Assume we have $n = k^2$ variables (k even), which are divided into k blocks of k variables each. The value of the function is 1 if there is at least one block with exactly two consecutive 1s in it, and it is 0 otherwise.

The block sensitivity of Rubinstein's function was shown to be equal to $n/2$ (hence, the certificate complexity and the decision-tree complexity is at least $n/2$) and the sensitivity is $\sqrt{n}$ and can be verified by a direct computation of $\hat{f}([n])$. Hence, for this function we can establish the relation:

$$bs(f) \geq s(f)^2/2$$

7 Hao's result

Aim: In this section, we expound Hao's result which shows that every $(2^{n-1}+1)$-vertex induced subgraph of the n-cube graph Q_n has maximum degree at least $\sqrt{n}$, hence proving the Sensitivity Conjecture.

7.1 Some Motivation

Rubinstein's example of quadratic discrepancy we had seen earlier is still the best known lower bound on $bs(f)$ in terms of $s(f)$. But no one had proven anything better than an exponential upper bound until Hao's result, from which it follows that: For all Boolean functions f,

$$bs(f) \leq 2s(f)^4$$

This bound is concrete, not asymptotic. It still leaves a gap between quadratic and quartic. It is however, the combination of two quadratic bounds. From (17), we conclude that:

$$bs(f) \leq 2\deg(f)^2 \tag{22}$$

The other is the conjecture,

$$\deg(f) \leq s(f)^2 \tag{23}$$

in Gotsman and Linial's paper, which is exactly what Hao proves.

Theorem 7.1. *(Hao) For all integers n greater than 1, let H be an arbitrary $(2^{n-1}+1)$-vertex induced subgraph of Q_n, then $\Delta(H) \geq \sqrt{n}$. Moreover, this inequality is tight when n is a perfect square.*

To prove this theorem, we will need a series of lemmas.

7.2 Cauchy's Interlace Theorem

Lemma 7.2. *All eigenvalues of a Hermitian matrix are real numbers.*

Proof. Let λ be an arbitrary eigenvalue of a Hermitian matrix A with corresponding eigenvector $\mathbf{x}$. Then:

$$A\mathbf{x} = \lambda\mathbf{x} \implies \mathbf{x}^T A\mathbf{x} = \mathbf{x}^T(\lambda\mathbf{x}) \implies \mathbf{x}^T A\mathbf{x} = \lambda\mathbf{x}^T\mathbf{x} = \lambda||\mathbf{x}||$$

Take the conjugate transpose on both sides to get (note A is Hermitian):

$$\mathbf{x}^T\bar{A}^T\mathbf{x} = \bar{\lambda}||\mathbf{x}|| \implies \mathbf{x}^T\bar{A}^T\mathbf{x} = \mathbf{x}^T A\mathbf{x} = \mathbf{x}^T\lambda\mathbf{x} = \lambda||\mathbf{x}||$$

Hence, $\lambda||\mathbf{x}|| = \bar{\lambda}||\mathbf{x}||$. We observe that $\mathbf{x}$ isn't the zero vector because it's an eigenvector. So, $\lambda = \bar{\lambda} \implies \lambda \in \mathbb{R}$. □

Now, if f and g are polynomials with real roots $r_1 \leq r_2 \leq \cdots \leq r_n$ and $s_1 \leq s_2 \leq \cdots s_{n-1}$, we say that f and g *interlace* if and only if $r_1 \leq s_1 \leq r_2 \leq \cdots \leq s_{n-1} \leq r_n$. We let the reader refer to [14] for the proof of this result.

Lemma 7.3. *The roots of polynomials f and g interlace if and only if the linear combinations $f + \alpha g$ have all real roots for all $\alpha \in \mathbf{R}$.*

Corollary 7.4. *If A is a Hermitian matrix, and B is a principal submatrix of A, then the eigenvalues of B interlace the eigenvalues of A.*

Proof. Simultaneously permuting rows and columns, if necessary, we may assume that the submatrix B occupies rows $2, \cdots, n$ and columns $2, \cdots, n$, so that A has the form:

$$A = \begin{pmatrix} B & \mathbf{c} \\ \mathbf{c}^* & d \end{pmatrix}$$

where $*$ signifies the conjugate transpose of a matrix. Choose $\alpha \in \mathbf{R}$, and consider the equation that follows from the linearity of the determinant:

$$\begin{vmatrix} B - xI & \mathbf{c} \\ \mathbf{c}^* & d - x + \alpha \end{vmatrix} = \begin{vmatrix} B - xI & \mathbf{c} \\ \mathbf{c}^* & d - x \end{vmatrix} + \begin{vmatrix} B - xI & \mathbf{c} \\ 0 & \alpha \end{vmatrix} \tag{24}$$

Now, we can conclude from Lemma 7.2 using the RHS of (24) that $|A - xI| + \alpha|B - xI|$ has all real roots for any α, so the eigenvalues interlace. □

This proves the following important result:

Lemma 7.5. *(Cauchy's Interlace Theorem) Let A be a symmetric $n \times n$ matrix, and let B be a $m \times m$ submatrix of A, for some $m < n$. If the eigenvalues of A are $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ and the eigenvalues of B are $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_m$, for all $1 \leq i \leq m$,*

$$\lambda_i \geq \mu_i \geq \lambda_{i+n-m}$$

7.3 Proving Hao's Theorem

Now, it can be easily shown that the degree $d(G)$ of a graph G is always at least as great as the largest eigenvalue λ_1 of the adjacency matrix A_G. So, let's take a m-vertex subgraph H of Q_n and let its top eigenvalue be $\lambda_1(H)$. Now, we know the entries of the adjacency matrix A_H are either 0 or 1. Hao's insight was to realise that if we flipped the signs of some 1s in A_H, the relation between the degree and the largest eigenvalue of the adjacency matrix still holds.

Lemma 7.6. *Suppose H is a m-vertex undirected graph, and A is a symmetric matrix with entries in $\{0, \pm 1\}$ and whose rows and columns are indexed by $V(H)$, the vertex set of H. Also, let $A_{u,v} = 0$ whenever $u, v \in V(H)$ are non-adjacent. Then,*

$$\Delta(H) \geq \lambda_1 := \lambda_1(A)$$

Proof. Let $\mathbf{v}$ be the eigenvector corresponding to λ_1. Then, $\lambda_1 \mathbf{v} = A\mathbf{v}$. We choose an index i that maximises the absolute value $|v_i|$. Then:

$$|\lambda_1 v_i| = |(A\mathbf{v})_i| = |\sum_j A_{i,j} v_j| \leq |\sum_j A_{i,j}| \cdot |v_i| \leq \sum_{(i,j) \in E(G)} |A_{i,j}| \cdot |v_i| \leq d(H)|v_i|$$

from which the result follows. □

Why is this lemma important? Observe that a key idea to this lemma was the introduction of -1 entries in the adjacency matrix A_H. We would like to know what happens if we try to use the original (unmodified) adjacency matrix in the lemma. One readily observes that though this matrix too has its top eigenvalue at least

$\sqrt{n}$, the interlacing bound is too lossy to prove this without modification. Take for example, the 3-dimensional cube Q_3, in which case, the eigenvalues of the unmodified adjacency matrix are $\{-3, -1, -1, -1, 1, 1, 1, 3\}$, giving the trivial bound $\lambda_1 \geq 0$.

Having proved this lemma, we now want to find conditions that force $\lambda_1 = \sqrt{n}$, especially when $m \geq \frac{N}{2} + 1$ where $N = |V(Q_n)| = 2^n$. Hao does this by making A sit inside a matrix A_N (whose construction will be discussed later) with at least $\frac{N}{2}$ eigenvalues with value $\sqrt{n}$.

For realising this, we construct A_{N-1} as a principal submatrix of A_N by deleting its last row and column. Since, A_N and A_{N-1} are both real and symmetric matrices, they have real eigenvalues and so we can order them as $\lambda_1, \cdots, \lambda_N$ and $\mu_1, \cdots, \mu_{N-1}$ in non-increasing order. From the property outlined before Lemma 7.3, we can see that their eigenvalues always interlace.

Now, we note that this process can be repeated, in the sense that, we can construct A_{N-2} by knocking off another row and its corresponding column and denoting its eigenvalues by ν_i, we can easily see that:

$$\mu_1 \geq \nu_1 \geq \mu_2 \geq \nu_2 \geq \mu_3 \cdots \implies \lambda_1 \geq \nu_1 \geq \lambda_3 \tag{25}$$

Repeating this procedure again gives us another matrix whose top eigenvalue is still at least as big as λ_4.

Generalising this for our purposes, if we repeat this process $\frac{N}{2} - 1 = 2^{n-1} - 1$ times inside A_N, the resulting matrix's leading eigenvalue is still at least as big as $\lambda_{\frac{N}{2}}(A_N)$, which we assign the value of $\sqrt{n}$. This means the size of the resultant (square) matrix will be $m = \frac{N}{2} + 1 = 2^{n-1} + 1$ and we conclude that

$$\lambda_1(A_N) \geq \lambda_1(A_m) \geq \lambda_{\frac{N}{2}}(A_N) = \sqrt{n}$$

which when used with Lemma 7.6 yields: $\boxed{d(H) \geq \sqrt{n}}$ which is nothing but Hao's result.

7.4 Constructing the Matrix

Now, as we've seen earlier, Hao showed that there exists a $N \times N$ matrix A_N with entries in $\{0, \pm 1\}$ whose non-zero entries correspond to the edges of the Boolean cube and such that all the N eigenvalues of A_N are $\pm\sqrt{n}$ and these eigenvalues sum up to zero, meaning that the trace is zero. This would then imply that $A_N^2 = nI$.

Also as we saw earlier, we would ideally want A_N to be the adjacency matrix of the n-cube but that doesn't work: that's because each (i, j) entry of the square of the unmodified induced adjacency matrix of Q_n counts all paths of length 2 from node i to node j and that number is nonzero.

This is where the trick of -1 on edges comes in. We *arrange* that every 4-cycle of the n-cube has exactly one edge with -1. Then, the pairs of paths from one corner to the opposite corner will always cancel, leaving $A_{i,j}^2 = 0$ whenever $i \neq j$. And,

$A^2_{i,i} = n$ because there are n ways to go out and come back along the same edge, always contributing $1 \cdot 1$ or $(-1) \cdot (-1)$ either way.

Subsequently, Hao defines the needed labelling exactly by the recursion:

$$A_2 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \text{ and } A_N = \begin{bmatrix} A_{\frac{N}{2}} & I \\ I & -A_{\frac{N}{2}} \end{bmatrix} \text{ for } (N > 2)$$

This construction puts a -1 sign on exactly one-fourth of the entries. We can now easily verify that $A^2_N = nI$ by induction.

8 Concluding Remarks

Aim: To provide a discussion of how the Sensitivity Conjecture is related to CREW PRAMs.

A PRAM (Parallel Random Access Machine) is the standard model for parallel computation. It consists of a set of processors $P_0, P_1, \cdots$ which communicate by means of cells $C_0, C_1, \cdots$, of shared memory. Now, each step of computation (of a function f) consists of three phases: read, compute and write phases. In the read phase, each processor may choose one cell to read from. In the compute phase, an arbitrary amount of local computation can take place. In the write phase, each processor may choose one cell to write into.

In the CREW (Concurrent Read Exclusive Write) PRAM variant, simultaneous read access is permitted, but not simultaneous write access. That is, several processors may read from the same location at the same time, but two or more processors may never attempt writing into the same location at the same time.

Now, a key result bounding the power of ideal CREW PRAMs is by Cook, Dwork and Reischuk [8] where they show that $\text{CREW}(f) = \Omega(\log(s(f)))$ is a lower bound on the number of steps required to compute a function f on a CREW PRAM. After this, Noam Nisan in [6] tweaked the definition of "sensitivity" and introduced the notion of "block sensitivity" to obtain: $\text{CREW}(f) = \Theta(\log(bs(f)))$. Now, the natural question was whether $\text{CREW}(f) = O(\log(s(f)))$? Proving this is nothing but the Sensitivity Conjecture.

References

[1] C. Gotsman, N. Linial. *The equivalence of two problems on the cube.* J. Combin. Theory Ser. A, **61** (1) (1992), 142–146.

[2] F. Chung, Z. Furedi, R. Graham, P. Seymour. *On induced subgraphs of the cube.* J. Comb. Theory, Ser. A, **49** (1) (1988), 180–187.

[3] N. Nisan, M. Szegedy. *On the degree of Boolean functions as real polynomials.* Comput. Complexity, **4** (1992), 462–467.

[4] Hao Huang. *Induced subgraphs of hypercubes and a proof of the Sensitivity Conjecture*. Annals of Mathematics, **190** (2019), 949-955.

[5] Ryan O'Donnell. *Analysis of Boolean Functions*, Cambridge University Press, 2014

[6] N. Nisan. *CREW PRAMs and decision trees*. Proc. 21st STOC, New York, NY, ACM Press (1989) 327–335.

[7] J. Kahn, G. Kalai, N. Linial. *The influence of variables on boolean functions*. FOCS (1988), 68-80.

[8] S. Cook, C. Dwork, R. Reischuk. *Upper and lower bounds for parallel random access machines without simultaneous writes*. Siam Journal on Computing, **15** (1) (1986), 87-97.

[9] H. Buhrman, R. de Wolf. *Complexity measures and decision tree complexity: a survey*, Theoretical Computer Science, **288** (2002), 21–43.

[10] P. Hatami, R. Kulkarni, D. Pankratov. *Variations on the Sensitivity Conjecture* Theory of Computing Library, Graduate Surveys **4** (2011), 1–27.

[11] M. Minsky, S. Papert. *Perceptrons*, MIT Press, Cambridge, 1988.

[12] N.I. Achieser. *Theory of approximation*, Dover Publications, 1992.

[13] Stasys Jukna. *Boolean Function Complexity*, Algorithms and Combinatorics **27**, Springer, 2012.

[14] Q.I. Rahman and G. Schmeisser. *Analytic Theory of Polynomials*, Oxford University Press, 2002.

General References

[15] Ken Regan. *Tools and Sensitivity*, accessed from https://rjlipton.wordpress.com/2019/07/12/tools-and-sensitivity/.

[16] Erica Klarreich, Quanta Magazine. *Decades-Old Computer Science Conjecture Solved In Two Pages*, July 2019.

Delhi Technological University
Shahbad Daulatpur, Main Bawana Road, New Delhi, Delhi, 110042
Corresponding email address: fromrohank07@gmail.com

arXiv:1912.00840v1 [eess.SY] 2 Dec 2019

OPTIMIZATION OF ELECTRIC VEHICLE CHARGING MANAGEMENT WITH DISTRIBUTIONAL LOCATIONAL MARGINAL PRICING

Abhishek Tyagi *
Department of Electrical Engineering
Delhi Technological University
New Delhi, DL 110042
abhishektyagi@dtu.ac.in

Ram Bhagat
Department of Electrical Engineering
Delhi Technological University
New Delhi, DL 110042
bhagat.ram@dtu.ac.in

ABSTRACT

Electric power generation, transmission, and distribution systems are attracting a large amount of interest from researchers with the development of the smart grid technologies. A smart grid aims at effective control and conditioning of the distribution of electricity. Pricing signal based distribution system are seen as one of the novel ways to achieve control in a smart grid. In our work, we propose to use a pricing signal modeled after the locational marginal price in the transmission system to locally provide price data to the users. The formulation and implementation of the distributional locational marginal price (DLMP) are achieved to develop a fair pricing model. The work is further practically implemented to a grid with Electric Vehicles in addition to the conventional load. The increasing popularity of EVs because of their ability to reduce greenhouse gas (GHG) emissions will pose a greater challenge in terms of demand on the power grid. We propose to use the DLMP modeling to alleviate congestion in the power grid and develop an optimal charging schedule for EVs.

***K*eywords** DLMP · Electric Vehicles · Open Source · Simulators · Benchmarks

1 Introduction

1.1 Smart Grids

The most important challenge of a power system is to operate itself optimally at the true demand-supply schedule in order to minimise its losses. Historically, power systems have utilised highly controlled supply systems usually from generation grids to manage deviations in the load[1]. However, owing to the recent awareness about climate and environment issues and the spurge in integration of information technology with power systems, "Smart Grid" technologies have started to enter the market. Electric power generation, transmission and distribution systems are attracting a large amount of interest from researchers as well as corporates alike with the development of the smart grid technologies. Smart Grids aim to utilise the extensive information from users to make power systems reliable, efficient and much more environment friendly[2].Smart Grids aim to act as inter-faces between generation system and consumers to coordinate the needs of the users with the capabilities of the generators and optimise the utilisation of resources[1, 2].

The envisioned smart grid aims to have the following features:

- *Integration of Renewable Resources:*Although alternate sources of energy have started penetrating in certain areas, complete implementation of these resources will require a more flexible management system[2]. Smart Grids utilise effective Demand-Side Management(DSM), more efficient use of transmission and intelligent storage to allow for the use of renewable generation in distribution systems. Also, the pricing infrastructure developed is such that it encourages the utilisation of renewable energy over conventional sources.[3]
- *Demand-Side Management*The problems associated with the integration of distributed renewable energy resources is dealt by a lot of researchers but an alternate approach to solving the problem is the involvement of

*

the consumer in the electricity market[4]. Pricing signals can be used as effective methods of control signals and responsive demand allows consumers to decrease their energy response on the basis of the availability of resources[5]. The first one refers to those changes applied by consumers to their electric load profile in response to energy market price signals for improving the economic efficiency of their energy consumption. This model adds to the efficiency of electricity markets by discouraging energy load when the real-time price is high and encouraging it when it is low. Further, this shall reduce the peak demand and make the electricity market more effective. [**?**, 6]

- *Congestion Management:* Smart Grids integrate congestion management from the transmission system to the distribution system as well. Solid State devices such as Flexible AC transmission (FACTS) controllers have been proposed for power flow control[4]. However, dynamic pricing signals can also be used to manage line flow limits during peak hours[7].

The last decade of the 20th century has seen growing concerns been raised regarding energy independence and climate change. Governments and other institutional organisations are working on developing policies to address this issue. The major focus is the consumption of fossil fuels and vehicular emission of green-house gases. In the U.S., the fuel economy standards on auto-mobiles have been set higher numerous times in the past ten years[8]. These concerns over carbon footprints have promoted the development of electric vehicles (EVs) as an alternative to petrol and diesel powered transport[9].Another major attraction for electric vehicles is the large scale integration of renewable energy into the power system. EVs are an attractive option to balance the uncertainties introduced in the system due to renewable energy[10].The number of electric vehicles on the road increased from zero in 1991 to two million in 2010 according to an Energy Foundation study[8]. Consequently, transportation economists have projected that EVs will comprise 64 to 86% of light vehicle sales in the US by 2030[11].

Although it has immense benefits, the large-scale integration of EVs into the power grid poses serious threats towards the stability of the grid. As suggested by Rahman and Shrestha, large scale deployment of EVs will eventually mean increased demand, especially at peak hours, impacting the reliability of the system. Demand patterns inducing flows that exceed the design limits to lead to congestion in the grid. Congestion from EVs can be observed at the medium voltage (MV) level, as demonstrated by a number of studies.[12, 13, 14] Hence, large scale uncoordinated charging of EVs can significantly affect network security[15]. Power quality degradation is also expected from the irregular demands][16]. These factors make it necessary to establish a coordinated standard for charging of EVs. The immense potential for smart grid technology lies in the various opportunities for control and coordination in power systems at the consumer end that have yet not been explored at all. Although a lot of research has been done recently in this regard, most of the research focuses on electronic controllers such as distribution class electronic switches and Solid State Transformers (SST)[17, 18].

In our work, we have developed a pricing infrastructure which can be used as a control feature of the smart grid. These pricing signals in turn will act as control signals to affect demand response, renewable energy penetration and congestion management. We propose to use a pricing signal modelled after the locational marginal price in transmission system to locally provide price data to the users. On a practical scale, we aim to utilise the model developed in Part-1 to develop a price based modelling approach to alleviate congestion induced by electric vehicles. Price based congestion management will ideally operate by the mechanism of generation re dispatch in response to the price signals.

2 Literature Review

Before we move towards developing the model to generate pricing signals, we shall look into the structure of the day-ahead and real-time electricity market and develop the theoretical concept of DLMP.

2.1 Electricity Market

The fact that electricity is a commodity that cannot be stored economically allows for the development of an electricity market place. To facilitate trade among power producers and consumers, an e-commerce marketplace called pool is developed and a framework is made to allow different types of contracts[19]. The market allows short-term trades that deliver electricity for the next day by utilising day-ahead pricing and intra-day markets that deliver electricity in the next 15 minutes[20].

2.2 Real Time Pricing

Throughout a day, there are certain hours that show low demand whereas we observe huge spikes in some places. Real Time price is an electricity rate that fluctuates according to the varying costs and amounts of generation and the demand

from the market. The main aim is to ensure that these real time prices reflect actual wholesale prices at different times of the day and the users are incentivised to shift loads to (o -peak) hours and hence ensure that the (peak-to-average ratio) is reduced[2, 21, 22]. An auction is carried out in the market every hour to calculate the market clearing price and consumption bids are taken from all end users.The concept of spot pricing for utilities was suggested by William Vickrey [23] in 1971.His idea was further developed by Bhon, Caramanis, and Schweppe [24], in 1984. They proposed a model which considered the marginal price of energy, losses, and congestion to determine real time prices. The dynamics of day ahead market is essential as it acts as the base price for the calculation of the real time prices. In a day-ahead market, different prices are quoted for each hour of the next day. The average is then taken to be the average of the day. These prices are considered to be the clearing price for the amount bid at that time. Any deviations from this bid can be bought or sold by the consumer at the real time market clearing price the next day[19, 20].

2.3 Locational Marginal Pricing

The Locational Marginal Price (LMP) represents the marginal cost to supply the next megawatt hour increment in power at the transmission bus without violating any security limits[4]. The LMP is calculated in the transmission system at every node and is calculated every 5 minutes on a real time market. According to Schweppe [25] the locational marginal price represents the truest cost of electricity. The major factors that affect the LMP at a bus are the generator bid prices, the amount of congestion experienced by the transmission system elements and the losses in the system[25]. Unlike the transmission system where nodal pricing is regularly used, distribution systems still make use of at pricing system leading to market inefficiency. Distribution Locational Marginal Price (DLMP) is defined as the real-time cost of providing one extra megawatt of energy at any point on the distribution system [2]. The implementation of such a model in the distribution system would lower total electricity costs for the end-user, mitigate the need to increase line load limits and also act as a perfect support mechanism to allow the utilisation of renewable generation[26, 4].

2.4 Factors Contributing To DLMP

The DLMP at any node is calculated by taking into account the transmission LMP and the distribution objectives and costs. Equation 1 is one of the formulations of DLMP.

$$DLMP = w_1 \sum \pi_i LMP_i + w_2 \sum (|I_k + \delta I_k|^2 - |I_k^2|) r_k + w_3 \sum P_\gamma + w_4 \sum \frac{|I_k|^2}{{I_k}^{\text{rating}}} \tag{1}$$

The DLMP can be broken down into four major components. These include:

- Marginal Energy Cost (MEC): This refers to the cost of energy at each substation and is calculated by the LM Pi at each substation and the contribution i of each substation.
- Marginal Loss Cost (MLC): The 2$^{\text{nd}}$nd term refers to the cost of I^2R losses in the branch k in the power system over an hour.
- Renewable Encouragement Factor: The 3$^{\text{rd}}$ term allows easy implementation of renewable energy and P denotes the cost of renewable generation over one hour.
- Marginal Congestion Cost (MCC): The 4$^{\text{th}}$ applies the penalty for the congestion in the line and penalises for high current loading

2.5 Demand Side Management in EV Charging

Demand Side Management is the modification of consumer demand through various methods such as financial implications to improve service and reduce consumption. Coordinated EV charging can be broadly classified on the basis of the decision making entity into centralized and distributed methods [27]. In case of centralized decision making, there exists large aggregators who manage multiple EVs. The EVs send their availability schedule and battery specifications to the aggregator who, on the basis of its profits [28, 29, 30], EV power costs [**?**], [31],[32] and reducing peak-to-average ratios [33]. On the other hand, in the distributed system, the flow of information is from the aggregator to the EVs who utilise their own understanding of the market to develop its charging plans. A mutual agreement needs to reached among all the parties as each party tries to minimise its individual costs [34] .Various optimisation techniques have been previously utilised in DSM literature to solve energy consumption scheduling problems. These include mixed-integer linear programming as done by Zhang et al. [35], quadratic programming for EV scheduling as done by Vandel et al. [36] and mixedinteger quadratic programming to allow for decentralised demand side management [37]. Most studies concentrate on mitigating the various impacts of EV charging load such as increases in peak load, power losses and decreased voltage.

2.6 Optimal EV Charging Management

Ruoyang Li et al [38] have theoretically developed an aggregator based EV charging management system to mitigate the negative impacts of EV loading on the grid. It considers the EV aggregators to be pro t seeking entities providing charging solutions to the EVs. The EV aggregators are price-takers and do not control the market. It thus proposes to forecast EV demands on the basis of all aggregators minimizing their individual costs whereas the conventional demand will be forecasted according to day-ahead prices. Figure 1 refers to the model developed to alleviate congestion caused



Figure 1: Congestion Alleviation from EVs using DLMPs

by EVs using pricing signals. The problem is solved from the point of view of the Distribution System Operator (DSO) to obtain the optimal solution for each EV aggregator. However, it can be easily proved that the efficient allocation of EV charging ${x_{i,t}}^*$ of the DSO problem is optimal for each EV aggregator under the DLMPs as well. As a corollary, the efficient allocation of the DSO problem ${x_{i,t}}^*, {c_{i,t}}^*$ can be achieved in a decentralized system under the DLMPs.

2.7 Karush-Kuhn-Tucker Conditions

The modelling and problem formulation utilises the Karush-Kuhn-Tucker (KKT) conditions (also known as the KuhnTucker conditions) to achieve optimization solution. First published in 1951 [41], the KKT conditions are first order necessary conditions for a solution in nonlinear programming to be optimal, provided that some regularity conditions are satisfied. KKT is essential in mathematical optimization as it generalises the method of Lagrange multi-pliers by allowing inequality constraints. In general, many optimization algorithms can be interpreted as methods for numerically solving the KKT system of equations. Through literature review, we can see the importance of LMP signals in distribution system and its use as a control signal in power systems. We have analysed the factors affecting the DLMP calculations and developed a formulation that takes care of all objectives and costs. Moreover, a detailed analysis of the demand side management techniques that have been previously used in EV charging specifically discussing different methods of decision making to decide EV charging schedule, has been done. A theoretical algorithm for coordinated EV charging from a centralized as well as decentralized perspective is discussed. In the subsequent chapters, we shall develop an optimisation problem to calculate DLMP and use it as a congestion alleviation tool.

3 Problem Formulation

In this section we'll formulate the problem that we have solved to determine the DLMP in a system. The cost objective function, constraints and it's relationship with DLMP is shown in the upcoming sections. We have utilised the AC-OPF (Optimal Power Flow) model to derive the DLMPs in Part I and DC-OPF model in Part II. The DCOPF is widely used and is considered to be sufficient for LMP calculation due to its computational efficiency and approximation accuracy[39].

3.1 Optimization Problem

In part I, AC-OPF optimization problem has an objective function which is to be minimized, whereas in part II, DC-OPF optimization- problem has an objective function which is to be maximized. These minimization and maximization are subject to constraints - both equality and inequality. To summarize we can write that,

$$\frac{min}{maxF(x)} \tag{2}$$

such that

$$G(x) = 0 \tag{3}$$

$$H(x) = 0 \tag{4}$$

where,
x: vector of dependent variables

3.2 Part 1

3.2.1 Objective Function

As discussed in Section 3.1, $F(x, u)$ is the objective function to be minimized. It is the total cost of energy to supply the demand in the system, given by,

$$F(x) = \sum_{i=1}^{G} LMP_i P_{Gi} + \sum_{i=1}^{DG} C_{DGi} P_{DGi} \tag{5}$$

where,
G : number of grids in the system
DG: number of renewable energy sources in the system
P_{Gi}: power generated by grid i
P_{DGi}: power generated by renewable generator i
LMP_i: price taken by grid i to deliver P_{Gi}
C_{DGi}: price taken by renewable generator to deliever P_{DGi}

3.2.2 Equality Constraints

As discussed in Section 3.1, the optimization problem has certain equality constraints, also known as Power Balance Constraints. They are obtained by balancing the real and reactive power at all the buses in the system, shown as below,

$$\sum_{i \epsilon G_k} P_{Gi} + \sum_{i \epsilon DG_k} P_{DGi} = P_k + f_{pk} \tag{6}$$

$$\sum_{i \epsilon G_k} Q_{Gi} + \sum_{i \epsilon DG_k} Q_{DGi} = Q_k + f_{qk} \tag{7}$$

where,
v_i: voltage at bus i i: angle at bus i G_k: set of grids connected to bus k DG_k: set of renewable generators connected to bus k P_k: real power load at bus k Q_k: reactive power load at bus k f_{pk}: real power owing out of bus k

$$f_{pk} = \sum_{i=1}^{N} v_k v_i (G_{ki} \cos(\theta_k - \theta_i) + B_{ki} (\sin(\theta_k - \theta_i))) \tag{8}$$

f_{qk}: reactive power owing out of bus k

$$f_{qk} = \sum_{i=1}^{N} v_k v_i (G_{ki} \sin(\theta_k - \theta_i) - B_{ki} (\cos(\theta_k - \theta_i))) \tag{9}$$

3.2.3 Inequality Constraints

The network inequality constraints for an AC-OPF optimization problem can be divided in general in three parts:
Line flow limits
There is a constraint in the power owing any two buses i and j i.e S_{ij}, given by

$$0 \leq S_{ij} \leq {S_{ij}}^{\max} \tag{10}$$

where,

$$S_{ij} = \sqrt{P_{ij}^2 + Q_{ij}^2} \tag{11}$$

$$P_{ij} = \frac{[v_i^2 - v_i v_j \cos(\theta_{ij})]r_{ij} + [v_i v_j \sin(\theta_{ij})]x_{ij}}{r_{ij}^2 + x_{ij}^2} \tag{12}$$

$$Q_{ij} = \frac{[v_i^2 - v_i v_j \cos(\theta_{ij})]x_{ij} - [v_i v_j \sin(\theta_{ij})]r_{ij}}{r_{ij}^2 + x_{ij}^2} \tag{13}$$

Voltage Limits
The voltage at all the buses can only vary between a minimum and a maximum value. Hence for any bus k,

$$v_k{}^{\min} \leq v_k \leq v_k{}^{\max} \tag{14}$$

Power Supply Limits
The power supplied by the grid and the renewable generators can also vary only in a given range. Hence for any grid or renewable generator i we can write,

$$P_i{}^{\min} \leq P_i \leq P_i{}^{\max} \tag{15}$$

3.2.4 Calculating DLMP

The optimization problem described in the Section 3.1 can be solved using the Lagrangian multipliers

$$L = F(x,u) + \sum_{k=1}^{N} \lambda_K G_K(x,u) + \sum_{l=1}^{M} \gamma_l H_l(x,u) \tag{16}$$

The DLMP at each bus k is simply the Lagrange multiplier associated with the real power balance constraint for that bus. [43]

$$DLMP_k = \lambda_k \tag{17}$$

3.3 Part II

3.3.1 Objective Function

As discussed in Section 3.1, F (x; u) is the objective function to be minimized. We are solving the problem from the DSO side and the objective is economic allocation for household demand and EV charging energy. Hence, the DSO objective is to maximise the social surplus, which can be mathematically presented as

$$maxF(x,u) \tag{18}$$

$$F(x,u) = \sum_{i \epsilon N} \sum_{t \epsilon T} \int_a^{c_{i,t}} P_{i,t}(\tau_i, t) d\tau_{i,t} - \sum_{t \epsilon T} P_{LMP_t} q_{g,t} \tag{19}$$

where,
N_c: Number of demand nodes
T: Number of hours for which system is analysed
$T_{i,t}$:Demand variable at time t and node i
$P_{i,t}(\tau_i, t)d\tau_{i,t}$: Benefits of using demand

$$P_{LMP_t}$$

: System LMP at time t for the node feeding the grid
$q_{g,t}$: Power supplied by the generator to the grid at t
$c_{i,t}$:Conventional household demand at time t at node i

The objective function consists of two terms. The first one represents the socialvalue of meeting the conventional demand, given by the area under the demand functions. The bene t of EV demand is not used as it is constant as long as the EV demand is met within the day [40]. The second term rep-resents the total cost of satisfying both the conventional

as well as the EV demand.
The optimization function can be further decomposed as follows:

$$F(x,u) = \sum_{i\epsilon N}\sum_{t\epsilon T}\int_{a}^{c_{i,t}} P_{i,t}(\tau_i,t)d\tau_{i,t} - \sum_{t\epsilon T} P_{LMP_t}\sum_{i\epsilon N_c}(c_{i,t}+x_{i,t}) \tag{20}$$

$$= \sum_{i\epsilon N}\sum_{t\epsilon T}\int_{a}^{c_{i,t}} P_{i,t}(\tau_i,t)d\tau_{i,t} - \sum_{t\epsilon T} P_{LMP_t}\sum_{i\epsilon N_c}(c_{i,t}) - \sum_{t\epsilon T} P_{LMP_t}\sum_{i\epsilon N_c} x_{i,t} \tag{21}$$

where,

$$\sum_{i\epsilon N}\sum_{t\epsilon T}\int_{a}^{c_{i,t}} P_{i,t}(\tau_i,t)d\tau_{i,t} - \sum_{t\epsilon T} P_{LMP_t}\sum_{i\epsilon N_c}(c_{i,t}) \tag{22}$$

is the social welfare corresponding to the conventional demand.

$$\sum_{t\epsilon T} P_{LMP_t}\sum_{i\epsilon N_c} x_{i,t} \tag{23}$$

is the EV charging cost.

3.3.2 Equality Constraints

Energy Balance Constraints
The total energy coming into the grid at time t should be same as the energy supplied by the generator. Net active power import/ export at time t for all nodes should be zero.

$$\sum_{i\epsilon N} r_{i,t} = 0 \forall t\epsilon T \tag{24}$$

where,

$$r_{i,t} = c_{i,t} + x_{i,t} \tag{25}$$

here,
$r_{i,t}$ is the net active power import/ export at time t at node i

Generation Node Balance Constraint
In this constraint the $r_{i,t}$ of the generator is made equal to the power supplied by the generator to the grid. Mathematically,

$$r_{g,t} + q_{g,t} = 0 \forall t\epsilon T \tag{26}$$

3.3.3 Inequality Constraints

Transmission Balance Constraint
There is a limit on the power that can flow in each line and hence the constraint is represented as,

$$-K_l \leq \sum_{i\epsilon N} D_{l,i} r_{i,t} \leq K_l \forall t\epsilon T, \forall l\epsilon L \tag{27}$$

where,
K_l: MVA capacity of line l
$D_{l,i}$: Power Transfer Distribution Factor (PTDF) coefficient of line l with respect to a unit injected at node i PTDF can be calculated as follows,

$$PTDF = diag(B) * A_r * ({A_r}^{\mathrm{T}} * diag(B) * A_r)^{-1} \tag{28}$$

here,

$$B = AX^1A^T \tag{29}$$

A: branch-to-node incidence matrix
X: diagonal reactance matrix
A_r: reduced incidence matrix

Charging Energy Limit Constraints
There is a limit on the maximum energy that a EV can take at one time from a charging station

$$0 \leq x_{i,t} \leq E_{i,t} \tag{30}$$

where,
$E_{i,t}$: EV battery size

Driving Requirement Constraints
The SOC (State of Charge) of EV batteries at time t for EV i is the sum of its initial SOC and total charging energy up to time period t-1 minus the total driving energy requirement up to time t.

$$S_{i,t^{`}} \leq S_{i,0} + \sum_{t^{`} \leq t-i} x_{i,t^{`}} - \sum t^{`} \leq td_{i,t^{`}} \leq S_{i,t^{`}} \forall i \epsilon N_c, \forall t \epsilon T - 1 \tag{31}$$

where,
$D_{i,t}$: distance travelled at time t by

3.3.4 Calculating DLMP

The DLMP for all the nodes in the system can be calculated by using the KKT multipliers found by solving the optimization problem defined in 3.4.1 using the constraints defined in 3.4.2 and 3.4.2.

$$DLMP = \begin{cases} -\mu^+_{i,t} + \mu^-_{i,t} - \sum_{t^{`} \leq t+1} K^+_{i,t^{`}} + \sum_{t^{`} \leq t+1} K^-_{i,t^{`}} & \forall_i \epsilon N_c . \forall_t \epsilon T - |T| \\ -\mu^+_{i,t} + \mu^-_{i,t} & \forall_t \epsilon |T| \end{cases}$$

We have theoretically formulated the optimization problem with the objective function and the constraints which were solved to gives us the DLMP at each bus. In the next chapter we will show how this theoretical approach can be modelled in software and can be utilised for calculation of DLMP in practical systems.

4 Modelling

4.1 Introduction

In the previous chapter in Section 3.2 we theoretically formed the optimization problem that needs to be solved to obtain the DLMP. This chapter deals with creating a model in MATLAB that solves the optimization problem and calculates the DLMP.

4.2 Non-Linear Programming Function MATLAB

We have used the MATLAB function $fmincon$ to solve the optimisation problem. $fmincon$ is a non-linear programming solver which gives the minimum of a function with respect to constraints.

$$minf(x) such that = \begin{cases} c(x) & \leq 0 \\ c_{eq}(x) & = 0 \\ A_{eq}.x = B_{eq} & \\ lb \leq x \leq ub & \end{cases}$$

Table 1: 12 Bus Distribution System Data

Voltage Base(kV)	V_B=2
MVA Base (MVA)	S_B=1
Feeder Impedance	z = 0.896+j0.7743
Feeder Length(mile)	L=0.5
Feeder Thermal Limit	I_{max} = 240A

4.3 Cost Function

We have used a separate function $myfun$ which returns the total cost of the energy summing the cost of energy supplied by each generator. We have used cost functions of two types in our analysis:

4.3.1 Polynomial Cost Function

The fuel cost function for grids or conventional thermal generators is represented by a quadratic polynomial function [44] given as,

$$C_i(P_{G_i}) = a_i + b_i P_{Gi} + c_i P_{G_i}^2 \tag{32}$$

4.3.2 Linear Cost Function

The cost for renewable energy is taken to be linear. It is assumed to contain the investment costs of equipments and the operation and management cost of generated energy. The cost function for renewable energy is hence given as

$$C_i(P_{GD_i}) = m_i P_{G_i} \tag{33}$$

Where, m_i is the slope of the assumed linear cost function.

4.4 Constraints

There are two types of constraints, equality and inequality. These constraints can be further divided as linear or non linear.

Linear Constraints
The voltage limits and power generation limits are linear.

Non-Linear Constraints
The power balance equations and Line flow limits are non linear constraints.

Having summarised the method in which we have modelled the DLMP in the software MATLAB, in the coming two chapters we'll consider a sample system of 12 buses, define its specifications and discuss the results of various cases on it.

5 System Specifications

In the previous section we have discussed modelling of the optimization problem in MATLAB. We shall utilise the model for two different systems. Firstly, a 12 bus system is studied to calculate DLMP across the system. Secondly, a part of the bus 4 distribution system of RBTS is used to demonstrate the efficacy of DLMP signals as an effective tool to alleviate congestion from EV demand. This chapter deals with the specifications of both the systems and the discusses the data selected for each case

5.1 Part-I

5.1.1 Elements and Circuit Diagram

The sample system we have considered is a 12 bus system as shown in Figure 2[4].

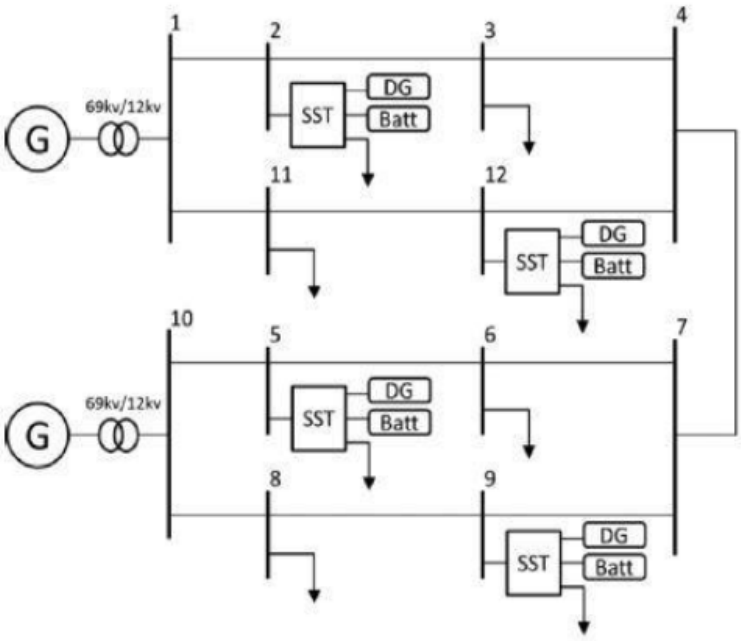


Figure 2: 12 Bus Distribution System

Table 2: Coefficients for Cost of Grid

Bus Number	$a_i(^{Rs.}/_{hr})$	$b_i(^{Rs.}/_{MWhr})$	$c_i(^{Rs.}/_{MW^2hr})$
1	0	3.00	0.00250
10	0	2.75	0.00625

Generators
It has two grids connected to bus 1 and 10. Four renewable generators are connected to buses 2, 5, 9 and 12. We have taken the renewable energy to be wind energy.

Loads
There are loads connected to buses 2, 3, 5, 6, 8, 9, 11 and 12, whose values are decided by the load profiles shown in Figure 6 in Section 6.2 for different cases.

Other Specifications of the Distribution System
Some other specifications of the system like base units and feeder impedance are shown in the Table 1.

5.1.2 Cost Funtions

As discussed in Section 4.3 the costs are of two types.
Grids
The cost of grids is taken to be quadratic polynomial and the coefficients a_i, b_i and c_i for the same are tabulated in Table 2[40]

Renewable Wind Generators
The cost of the renewable wind generators is taken as linear passing through origin and the slope for the same is tabulated in the Table 3. [40]

Penalty in Real Time
In the day ahead market power from the grids as well as the wind generators are decided based on the forecasted load.

Table 3: Coefficients for Cost of Renewable Generators

Bus Number	$m_i(^{Rs.}/_{MWhr})$
2	3.25
5	3.5
9	3.25
12	3.5

Table 4: Connection Line Types

Line Type	Line Length	Line Number
1	0.6	2 6 10 14
2	0.75	1 4 7 9 12
3	0.8	3 5 8 11 13

But since in the real time market the actual load differs from the forecasted load, some extra power is drawn from either the grid or the wind generators. The grid in real time imposes a penalty i.e. charges a high price if more than scheduled load is drawn. In such a case getting power from the wind generators become cheaper. But there is a limit on the power generated by the wind generator (Section 3.4.3), so when its maximum is reached, the system is forced to draw power from the generator leading to an increase in the cost.

5.2 Part-II

5.2.1 System Outline

A sub part of the Bus 4 distribution network of the RBTS system is used as a case study to test the model.

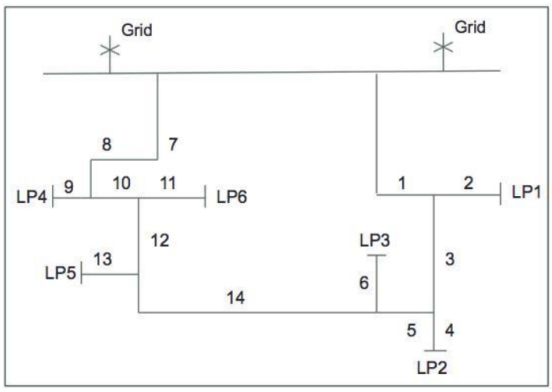


Figure 3: Single Line Diagram of test system

Figure 3 [40] shows the single line diagram of the test system. The system has 2 supply points, 6 load points and 2 EV aggregators (each consisting of 25 EVs under its ambit). The electric distribution systems of RBTS represent typical distribution networks and follow the general principles and practices regarding topology and ratings. Table 4 lists the different types of lines along with respective line lengths.

Customer Data
The customer data consists of peak and average loads for each load point. The inverse demand function at each bus is assumed to be linear with a price elasticity of 0.1 as shown in Figure 4. This level of demand price elasticity is consistent with empirical studies.

EV
Table 6 summarises the EV data. The battery size of each EV varies according to EV driving requirements. It is assumed that the maximum charging power is 1.15 kW (based on a 5 A, 230 V connection). The energy consumption while

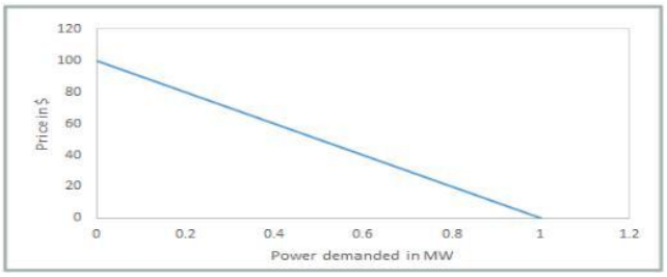


Figure 4: Demand Curve of each Load

Table 5: Customer Level Data

Load Points	Average Load/LP (MW)	Peak Load/LP(MW)
1,3,4,5	1.0	1.63
2,6	1.5	2.445

driving is kept at 0.15 kWh/km. The EV battery can fluctuate between a maximum SOC of 85% and a minimum of 20%. Each EV has an individual initial SOC that is chosen according to its driving schedule.

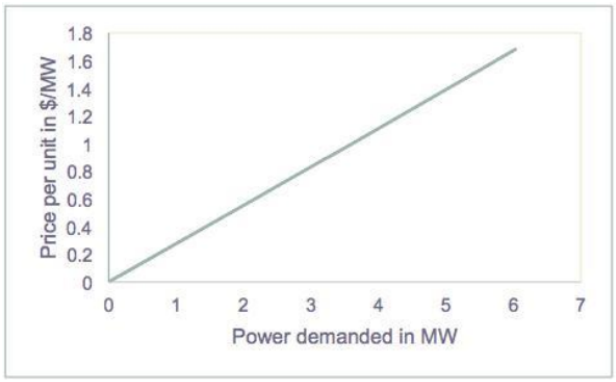


Figure 5: Generator Cost Curve

Generator Cost Function
The generator cost function is defined as a quadratic polynomial represented by the equation

$$P = 0.275Q + 0.0006250Q^2 \tag{34}$$

where Q is the energy demand and P is the cost Figure 5 shows the graph of the generator cost with the net energy demand.

In this section we have defined the system specifications, limits and cost functions. Using these values in the model (Section 4) we'll run different cases which will be discussed in the next Section.

6 Results and Discussions

This section is divided in two parts - Part I and Part II. For each part we first present the data that is used as input for the system and then we discuss the observations and conclusions for different scenarios.

6.1 Part-I

In this section we shall discuss the following scenarios:

- Day ahead market

Table 6: EV Data Summary

EV Parameter	Value
EV Battery Size	25 kWh
Energy Consumption of Driving	0.15kWh/km
Minimum SOC	20%
Maximum SOC	80%

- Real time market
- Real time market with responsive loads
- Real time market with congestion

6.1.1 Data Used

To get the day-ahead pricing signal, we shall use the forecast of the load for 24 hours as shown in Figure 6 [48] and the forecast of the wind energy production, as shown in

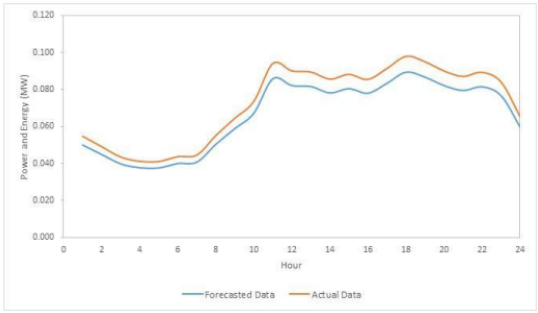


Figure 6: Load Profile-Forecast and Actual

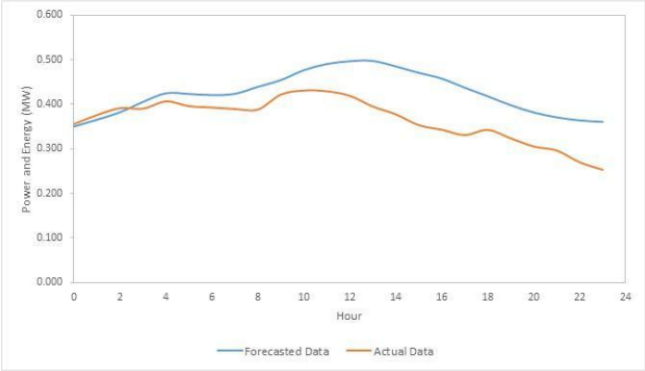


Figure 7: Wind Power Generation Profile-Forecast and Actual

6.1.2 Results

Day ahead market As discussed in Section 2.1.2, the bidding for the day ahead market happens 24 hours prior to the day when the actual requirement occurs. Since the actual load profile and wind power generation data is not available, we use the forecasted data to predict the production bills by all renewable generators as well as the grid.

Figure 8 shows a comparison of the Total Load, total contribution from both the grids and the renewable sources and the value of the total cost function over the period of 24 hours. From Figure 8, it can be observed that as the total load in the system increases, the total cost as well as the contribution of wind energy increases. However the contribution of

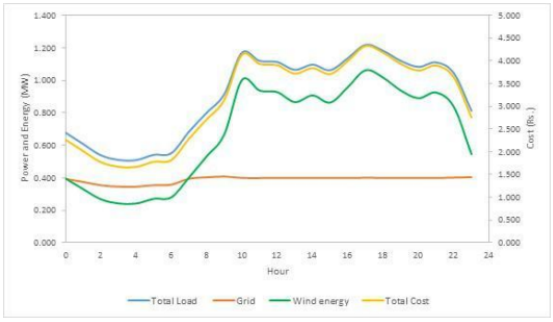


Figure 8: Day Ahead: Comparison of load, generation cost over 24 hours

grid remains almost the same. This is beneficial for the system as the renewable sources are able to take care of load variations and we do not have to retrieve extra energy from the grid. The grid supplies up to a predefined rate at a price which is cheaper than renewable sources as discussed in 5.2.2. After that level, it is cheaper to take from the DGs, in order to promote these sources of energy.

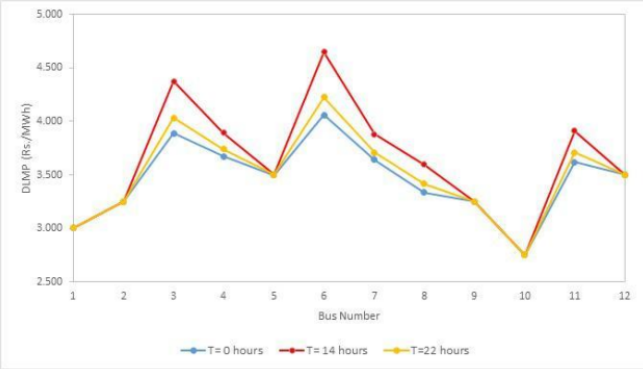


Figure 9: Day Ahead Market: Comparison of DLMP at different times

Figure 9 shows the variation in DLMP at different hours of the day for all the buses. It is very interesting to see that the DLMP is highest at t=14 hours which is a peak load hour and lowest at t=0 hours which is an off peak hour. Hence, it will provide an added incentive to the end-user to shift his loads to o-peak hours. The DLMP at buses 1, 2, 5, 9, 10

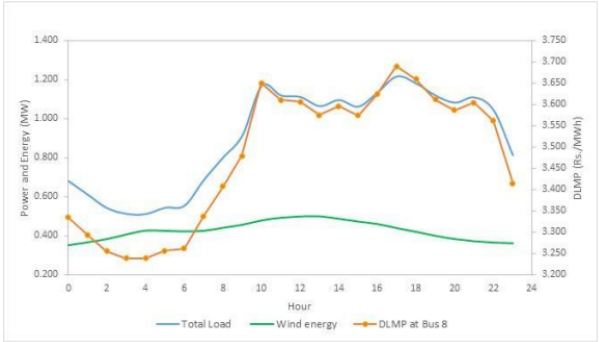


Figure 10: Day Ahead: Variation in DLMP with load and wind generation

and 12 are lower as compared to the other buses. This is because at these buses we either have a grid or a wind power generator. Since in supplying the loads at these buses the line loss component of DLMP doesn't come into picture, the DLMP is significantly less. Figure 10 is highlights the variations in DLMP with the changing load and wind generation

pro les at bus 8. An increase in the DLMP is seen with the increasing load as is expected. Further, when the generation is less, an increase in the DLMP is observed.

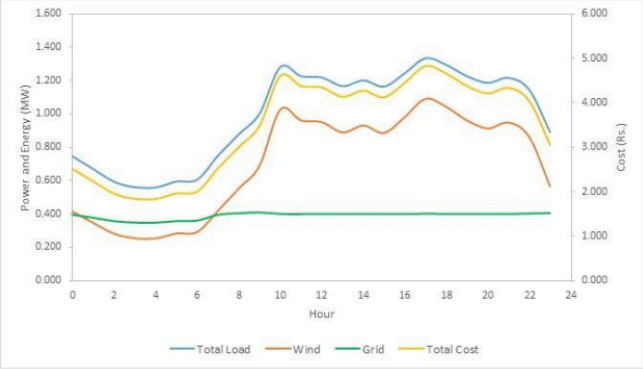


Figure 11: Real Time: Comparison of the load generation cost over 24 hours

Real Time Market

In the real-time market, the bidding is usually done 15 minutes before the energy is required at the load. At this moment, we have actual data relating to the load requirements at all buses and the generation by the renewable sources and hence that is used. The load usually shows a positive diversion from the forecast and is higher. It is this difference in load that has to be auctioned to the production companies. Also, the actual generation by renewable sources might be less as compared to the forecast and in case if the day-ahead bid amount is not produced on one of the buses, the concerned generator has to buy it from other sources.

Figure 11 shows a comparison of the Total Load, total contribution from both the grid and the renewable sources and

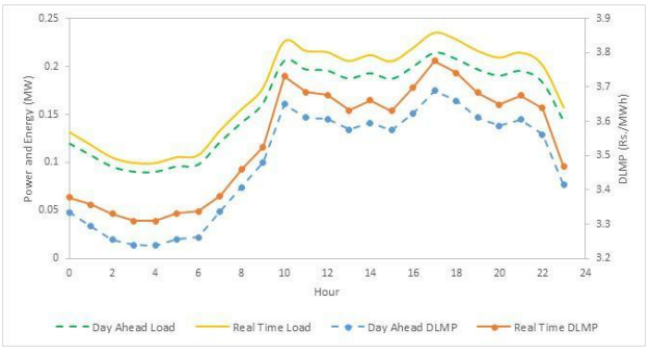


Figure 12: Comparison of DLMP and Load for Real Time and Day Ahead

the value of the total cost function over the period of 24 hours in the real time market. The trends followed are similar to Figure 8.

As the load requirements in the real time increase as compared to the day ahead case, the DLMP signals also show a spurge. This is seen in Figure 12. Hence, an added penalty is applied to the loads that divert from the initially bid amount making the process of buying electricity in the real time market expensive. This makes economic sense as well since if there is a load in urgent need of electricity, the price charged can be increased. It is also seen that the change in DLMP is proportional to the change in the load. If major diversions from the initial bid are made, the DLMP will shoot up.Also, as in the day ahead case, the DLMP shows spikes in value when the load increases on the corresponding bus. This behaviour is seen in Figure 13 for the case of bus 8 and can be a major deterrent for peak loading by different consumers.

Real time market with responsive loads

The results for the real time indicate that all the buses are at their maximum loads with Bus 6 having the highest DLMP of 5.21 as well as the maximum load of 0.23 MW. In a responsive system, the load should ideally adjust its consumption with respect to the price signal. The load at Bus 6 is made responsive to see the effect of demand response. It can adjust its load from the maximum value of 0.23MW to 0 MW.

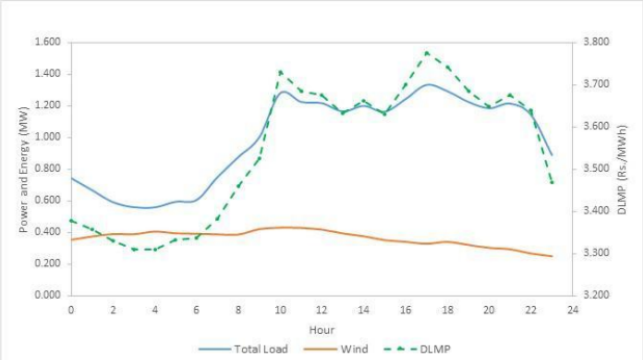


Figure 13: Real Time: Variation in DLMP with load and wind generation

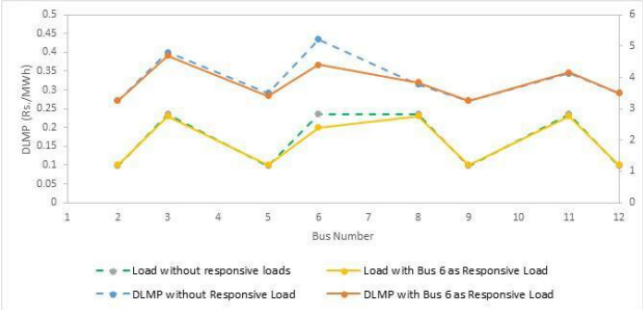


Figure 14: Addition of Responsive Load: Effect on DLMP and Load

The results of the test shows clear instance of demand response. When the load at Bus 6 is made responsive, it responds by reducing its consumption to 0.20 MW. The DLMP at Bus 6 also goes down to 4.40 from 5.21 indicating adequate amount of responsivity in the demand. Hence, we can reduce the demand in peak hours as well (peak-to-average ratio) by implementing demand response in smart grid systems.

Real time market with congestion
In the real time, the power owing in branch between buses 11-12 was seen to be the maximum. Line limits on this branch are changed to 0.1 MW from 3.2 MW to induce congestion on this line. The AC OPF solution illustrates that the injection from the generator at bus 12 decreases as it will no longer be able to supply to bus 11. As a result, the DLMP at Bus 11 sees a spike and goes upto a maximum value of 5.94.Practically this is because the load at Bus 11 cannot take power either from Bus 12 due to the line congestion limit and will have to look upto Bus 0 or Bus 2 for its load requirements. Since, this is an added pressure and will also incur additional losses, DLMP shows a sharp increase in value.

Driving Data
The driving data used in the model is from the Danish National Travel Survey [13]. The Danish driving data were chosen because the driving behavior in Denmark could be representative of the EV users driving pattern. In Denmark, the average driving distance is about 40 km per day. Customers who need to drive a longer distance might not choose to use EVs. To get the profile of driving pattern for the EV customers in our model we assumed a total of 50 EVs. Then using Table A.11 we get the number of EVs travelling a distance of 0-10km, 10-20km and so on every day. Using the EV availability curve from [40], we divide the distance of each EV travels equally among the hours that it is available for travelling. We eventually combined the first 25 EVs under one EV aggregator and the last 25 EVs under the 2nd EV aggregtor.Figure 17 shows the sample EV availability for an EV. At the times which are marked black the EV is travelling and hence cannot be charged but for the times for which it is shown white it is available for charging.

6.1.3 Case-I

Case I represents the most basic case with 2 EV aggregators and the system at 100% EV penetration. Figure 17 and Figure 18 shows the availability of both EV aggregators for this case. Figure 19 depicts the distance travelled by each

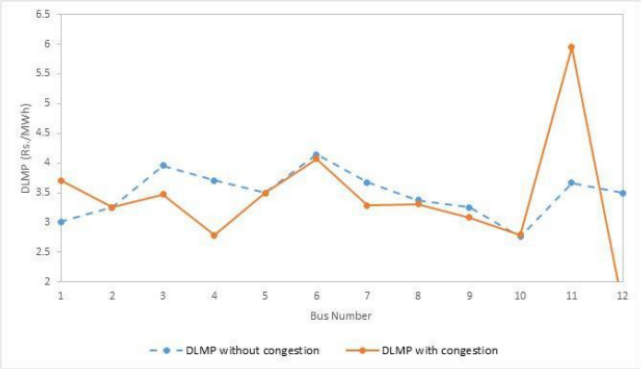


Figure 15: Congestion: Effect on DLMP

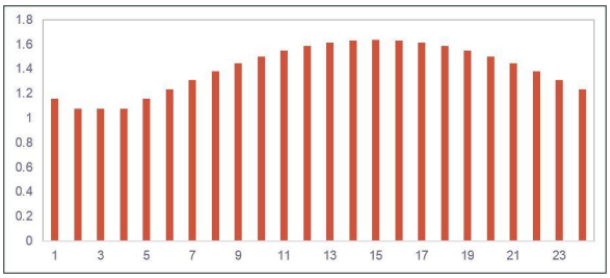


Figure 16: Conventional Customers Demand Curve



Figure 17: Availability curve for EV1

EV at every hour when it is on duty. The results of Case Study I are shown in Figure 20-Figure 22. The day begins with



Figure 18: Availability curve for EV2

an initial EV charge of 25%. Figs 6.16 and 6.17 illustrate the use of DLMPs in reducing the impact of EV loads on the grid at any particular hour. Comparing Figure 21 with Figure 20, we see that EV loads are spread out across multiple hours according to the LMPs at each hour instead of charging at any one hour. This is because charging at one hour would lead to a serious spike in the value of DLMP at that hour and therefore leading to instability in the system. As seen, the EVs charge at times with minimum conventional demand so as to minimize their charging costs and maximize social welfare.

6.1.4 Case-II

Case II represents a case with 2 EV aggregators and the system at 100% EV penetration. Figure 17 and Figure 18 shows the availability of both EV aggregators for this case. In both cases the availability is such that it necessitates that the EVs charge during the high price hours. Figure 19 depicts the distance travelled by each EV at every hour when it is on duty. The driving schedule of each EV.The results of Case Study II are shown in Figure 20-Figure 22. The difference in Case II with respect to Case I is evident from Figure 26 and Figure 27. The EV availability schedule is such that the

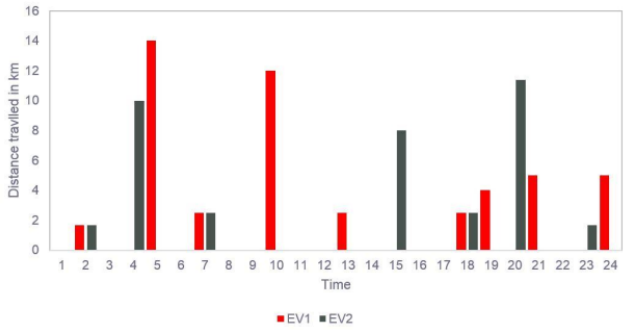


Figure 19: Distance travelled by both EVs across 24 hours

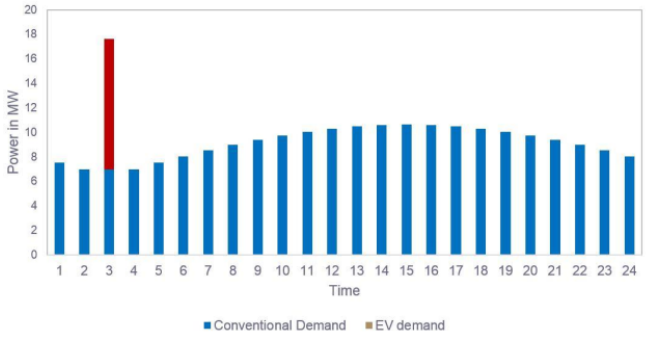


Figure 20: Line 1 loading without DLMPs in Case I

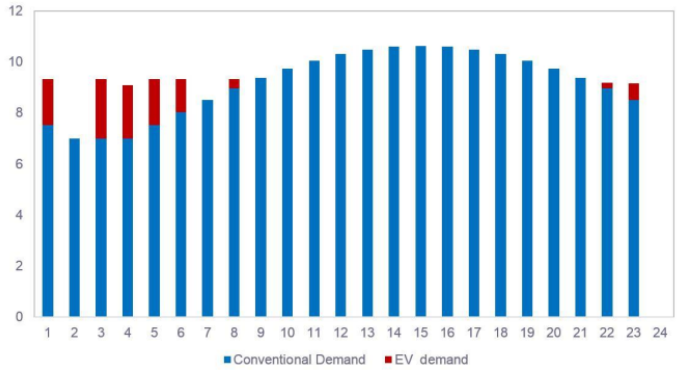


Figure 21: Line 1 loading with DLMPs in Case I

EVs are forced to charge in hours 12, 13 and 14 when the LMPs are generally high. However, even in this case the charging is balanced rather than concentrated at one particular hour.

6.1.5 Case-III

Case III represents a system which has 500% EV penetration in terms of the net EV demand. **??** and Figure 29 illustrate the results for the case. The essential point in this case is the EV loading after DLMPs are used in the system. We see an optimized distribution of EV loading across the 24 hours such as the hours that have low LMP see higher EV load as compared to hours that have high LMP. This is almost constant across time except for hours when the EVs are driving. This ultimately leads to an approximately equitable demand at each hour for the system and hence means a general stability of the system.

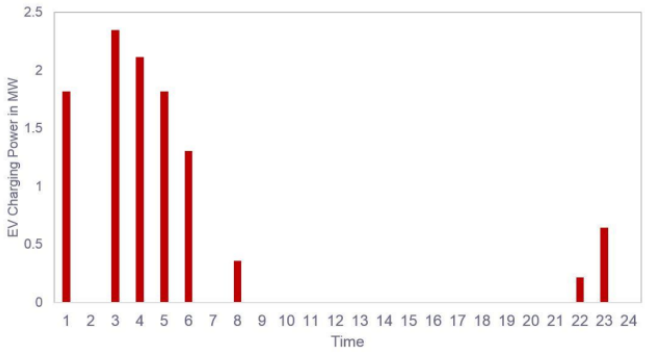


Figure 22: Total EV loads with DLMPs in Case I



Figure 23: Availability curve for EV1 for Case II



Figure 24: Availability curve for EV2 for Case II

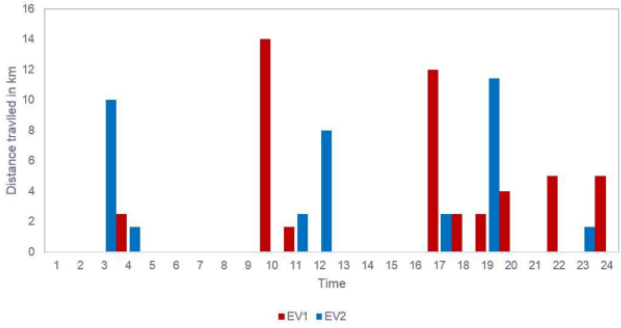


Figure 25: Distance travelled by both EVs across 24 hours for Case II

6.1.6 Case-IV

Case IV consists of the same system as in Case I but with congestion induced in the line 14 by reducing the limits of the line. Figure 30 and Figure 31 represent the results achieved. As a result of the congestion in the system, there is a visible change in the power used to charge the EVs at time 4 when the systems meets the limits.

Fig 6.27 depicts a comparison between the EV demand with and without congestion. The induced congestion is at time 4.As a result this EV charging at time 4 is reduced and in turn compensated at other times of the day.

Figure 32 represents the variation of DLMP and LMP with time at node 11. The DLMP follows the LMP at all times except at time 4 when congestion is induced into the system. This is because the system is lossless system and the distribution price is same as the transmission price. However, at time t=4, congestion in the adjoining line causes a spurt in price of demand indicating a model that discourages users to charge at times when the load is high and shift load accordingly to other hours.

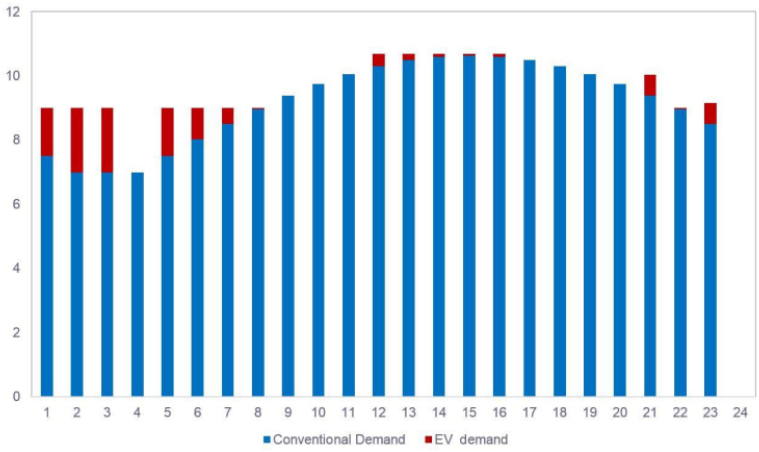


Figure 26: Line 1 loading with DLMPs in Case II

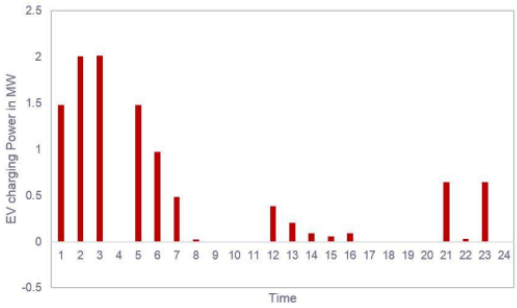


Figure 27: Total EV loads with DLMPs in Case II

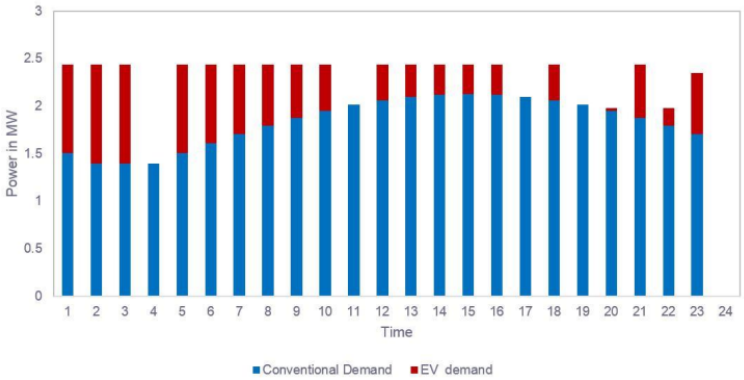


Figure 28: Line 1 loading with DLMPs in Case III

7 Conclusion

This work has two main conclusions. One of the main conclusion of this work is the development of a pricing signal modelled on the lines of Locational Marginal Pricing (LMP) in distribution systems as a control signal. A multi objective optimisation function has been developed for the calculation of the pricing signal. The implementation of the optimisation was developed in MATLAB and a robust model has been developed for the same. Various control objectives such as reduction in cost of energy, reduction of demand in peak hours, improvement of (peak-to-average ratio) and controlled flow in selected lines was successfully achieved. Secondly, a real-life practical implementation of the model is developed for a system with Electric Vehicle based loads in addition to the conventional household demand.

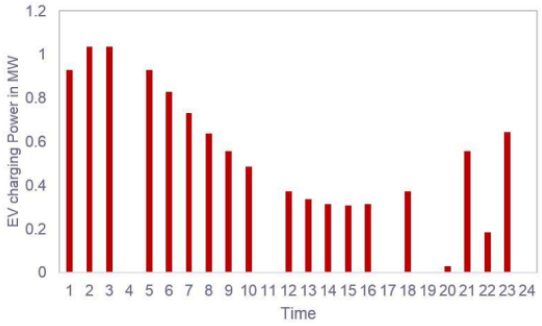


Figure 29: Total EV loads with DLMPs in Case III

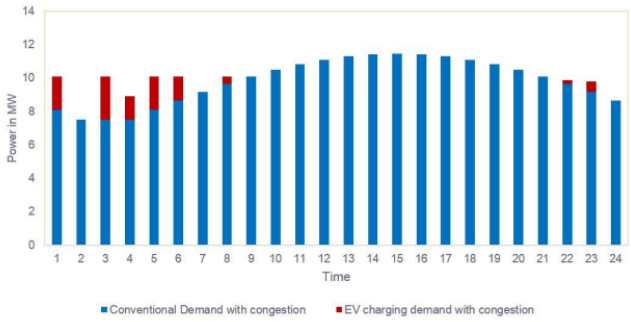


Figure 30: Line 1 loading with DLMPs in Case IV

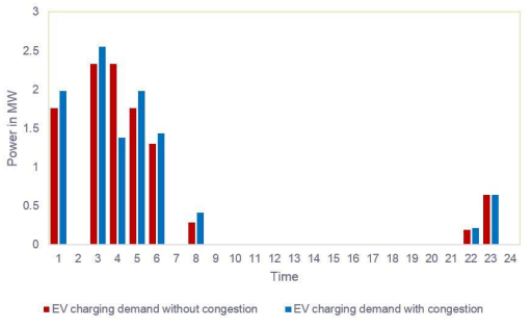


Figure 31: Comparison between EV loads with and without congestion

The model is developed with the aim to mitigate the risks that the grid might face due to large scale integration of EVs. Again,a non-linear constrained optimization schedule is developed and tested on a part of the Bus 4 RBTS system. The model is seen to have successfully alleviated congestion and can be used as a tool for demand side management and EV charge scheduling.

8 Future Work

The demand response shows interesting results and can be practically used to reduce peak power demand. The modelling can be used as a suitable benchmark for the study and development of distribution systems in the future smart grid. The future work requires further development of the model, introduction of other control objectives and testing the model on real-time cases. A well developed pricing infrastructure has the potential to work as a competent control mechanism for the future smart grid.
For the part II of the model, immense scope lies in adding additional feature of spatial optimization across different

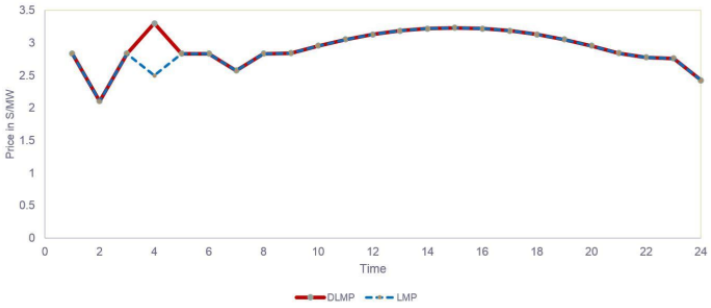


Figure 32: DLMPs at node 11 in Case III

charging stations. We can modify the system to allow EV users to drive to a different station to specifically charge. Further, a higher degree of precision can be achieved by including practical features such as driving distance, opportunity cost of driving time and spatial difference in EV charging costs. We can also expand the system to a multitude of EV aggregators. Further, the system can be expanded to a micro level by expanding the optimization to individual EV aggregators as well.

References

[1] C. W. Potter, A. Archambault, and K. Westrick. Building a smarter smart grid through better renewable energy information. In *2009 IEEE/PES Power Systems Conference and Exposition*, pages 1–5, March 2009.

[2] F. Sahriatzadeh, P. Nirbhavane, and A. K. Srivastava. Locational marginal price for distribution system considering demand response. In *2012 North American Power Symposium (NAPS)*, pages 1–5, Sep. 2012.

[3] G. T. Heydt, B. H. Chowdhury, M. L. Crow, D. Haughton, B. D. Kiefer, F. Meng, and B. R. Sathyanarayana. Pricing and control in the next generation power distribution system. *IEEE Transactions on Smart Grid*, 3(2):907–914, June 2012.

[4] C. Cecati, C. Citro, and P. Siano. Combined operations of renewable energy systems and responsive demand in a smart grid. *IEEE Transactions on Sustainable Energy*, 2(4):468–476, Oct 2011.

[5] A. J. Roscoe and G. Ault. Supporting high penetrations of renewable generation via implementation of real-time electricity pricing and demand response. *IET Renewable Power Generation*, 4(4):369–382, July 2010.

[6] C. Ibars, M. Navarro, and L. Giupponi. Distributed demand management in smart grid with a congestion game. In *2010 First IEEE International Conference on Smart Grid Communications*, pages 495–500, Oct 2010.

[7] P. Zhang, K. Qian, C. Zhou, B. G. Stewart, and D. M. Hepburn. A methodology for optimization of power systems demand due to electric vehicle charging load. *IEEE Transactions on Power Systems*, 27(3):1628–1636, Aug 2012.

[8] Q. Wu, A. H. Nielsen, J. Ostergaard, S. T. Cha, F. Marra, Y. Chen, and C. Træholt. Driving pattern analysis for electric vehicle (ev) grid integration study. In *2010 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT Europe)*, pages 1–6, Oct 2010.

[9] Thomas A Becker, Ikhlaq Sidhu, and Burghardt Tenderich. Electric vehicles in the united states: a new model with forecasts to 2030. *Center for Entrepreneurship and Technology, University of California, Berkeley*, 24, 2009.

[10] K. Clement-Nyns, E. Haesen, and J. Driesen. The impact of charging plug-in hybrid electric vehicles on a residential distribution grid. *IEEE Transactions on Power Systems*, 25(1):371–380, Feb 2010.

[11] K. J. Dyke, N. Schofield, and M. Barnes. The impact of transport electrification on electrical networks. *IEEE Transactions on Industrial Electronics*, 57(12):3917–3926, Dec 2010.

[12] L. Pieltain Fernández, T. Gomez San Roman, R. Cossent, C. Mateo Domingo, and P. Frías. Assessment of the impact of plug-in electric vehicles on distribution networks. *IEEE Transactions on Power Systems*, 26(1):206–213, Feb 2011.

[13] M. K. Gray and W. G. Morsi. Power quality assessment in distribution systems embedded with plug-in hybrid and battery electric vehicles. *IEEE Transactions on Power Systems*, 30(2):663–671, March 2015.

[14] S. Bhattacharya, T. Zhao, G. Wang, S. Dutta, S. Baek, Y. Du, B. Parkhideh, X. Zhou, and A. Q. Huang. Design and development of generation-i silicon based solid state transformer. In *2010 Twenty-Fifth Annual IEEE Applied Power Electronics Conference and Exposition (APEC)*, pages 1666–1673, Feb 2010.

[15] A. Mohsenian-Rad and A. Leon-Garcia. Optimal residential load control with price prediction in real-time electricity pricing environments. *IEEE Transactions on Smart Grid*, 1(2):120–133, Sep. 2010.

[16] Ronald Huisman, Christian Huurman, and Ronald Mahieu. Hourly electricity prices in day-ahead markets. *Energy Economics*, 29(2):240–248, 2007.

[17] F. J. Nogales, J. Contreras, A. J. Conejo, and R. Espinola. Forecasting next-day electricity prices by time series models. *IEEE Transactions on Power Systems*, 17(2):342–348, May 2002.

[18] M. Fahrioglu and F. L. Alvarado. Designing incentive compatible contracts for effective demand management. *IEEE Transactions on Power Systems*, 15(4):1255–1260, Nov 2000.

[19] William Vickrey. Responsive pricing of public utility services. *The Bell Journal of Economics and Management Science*, 2(1):337–346, 1971.

[20] Roger E Bohn, Michael C Caramanis, and Fred C Schweppe. Optimal pricing in electrical networks over space and time. *The Rand Journal of Economics*, pages 360–376, 1984.

[21] Jim Frame. Locational marginal pricing. In *2001 IEEE Power Engineering Society Winter Meeting. Conference Proceedings (Cat. No. 01CH37194)*, volume 1, pages 377–vol. IEEE, 2001.

[22] Stephen P. Holland and Erin T. Mansur. The short-run effects of time-varying prices in competitive electricity markets. *The Energy Journal*, 27(4):127–155, 2006.

[23] James JQ Yu, Junhao Lin, Albert Lam, and Victor OK Li. Coordinated electric vehicle charging control with aggregator power trading and indirect load control. *arXiv preprint arXiv:1508.00663*, 2015.

[24] S. Han, S. Han, and K. Sezaki. Development of an optimal vehicle-to-grid aggregator for frequency regulation. *IEEE Transactions on Smart Grid*, 1(1):65–72, June 2010.

[25] E. Sortomme and M. A. El-Sharkawi. Optimal charging strategies for unidirectional vehicle-to-grid. *IEEE Transactions on Smart Grid*, 2(1):131–138, March 2011.

[26] M. Ş. Kuran, A. Carneiro Viana, L. Iannone, D. Kofman, G. Mermoud, and J. P. Vasseur. A smart parking lot management system for scheduling the recharging of electric vehicles. *IEEE Transactions on Smart Grid*, 6(6):2942–2953, Nov 2015.

[27] M. H. K. Tushar, C. Assi, M. Maier, and M. F. Uddin. Smart microgrids: Optimal joint scheduling for electric vehicles and home appliances. *IEEE Transactions on Smart Grid*, 5(1):239–250, Jan 2014.

[28] D. Wu, D. C. Aliprantis, and L. Ying. Load scheduling and dispatch for aggregators of plug-in electric vehicles. *IEEE Transactions on Smart Grid*, 3(1):368–376, March 2012.

[29] D. T. Nguyen and L. B. Le. Joint optimization of electric vehicle and home energy scheduling considering user comfort preference. *IEEE Transactions on Smart Grid*, 5(1):188–199, Jan 2014.

[30] J. de Hoog, T. Alpcan, M. Brazil, D. A. Thomas, and I. Mareels. Optimal charging of electric vehicles taking distribution network constraints into account. *IEEE Transactions on Power Systems*, 30(1):365–375, Jan 2015.

[31] Di Zhang, Lazaros G Papageorgiou, Nouri J Samsatli, and Nilay Shah. Optimal scheduling of smart homes energy consumption with microgrid. In *PROCEEDINGS OF THE FIRST INTERNATIONAL CONFERENCE ON SMART GRIDS, GREEN COMMUNICATIONS AND IT ENERGY-AWARE TECHNOLOGIES (ENERGY 2011)*, pages 70–75. IARIA XPS PRESS, 2011.

[32] Y. Mou, H. Xing, Z. Lin, and M. Fu. Decentralized optimal demand-side management for phev charging in a smart grid. *IEEE Transactions on Smart Grid*, 6(2):726–736, March 2015.

[33] Sarvapali D Ramchurn, Perukrishnen Vytelingum, Alex Rogers, and Nick Jennings. Agent-based control for decentralised demand side management in the smart grid. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 5–12. International Foundation for Autonomous Agents and Multiagent Systems, 2011.

[34] R. Li, Q. Wu, and S. S. Oren. Distribution locational marginal pricing for optimal electric vehicle charging management. *IEEE Transactions on Power Systems*, 29(1):203–211, Jan 2014.

[35] H. W. Kuhn and A. W. Tucker. Nonlinear programming. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 481–492, Berkeley, Calif., 1951. University of California Press.

[36] F. Li and R. Bo. Dcopf-based lmp simulation: Algorithm, comparison with acopf, and sensitivity. *IEEE Transactions on Power Systems*, 22(4):1475–1485, Nov 2007.

[37] Haifeng Liu, Leigh Tesfatsion, and Ali A Chowdhury. Derivation of locational marginal prices for restructured wholesale power markets. *Journal of Energy Markets*, 2(1):3, 2009.

[38] Saoussen Brini, Hsan Hadj Abdallah, and Abderrazak Ouali. Economic dispatch for power system included wind and solar thermal energy. *Leonardo Journal of Sciences*, 14(2009):204–220, 2009.

[39] N. Augustine, S. Suresh, P. Moghe, and K. Sheikh. Economic dispatch for a microgrid considering renewable energy cost functions. In *2012 IEEE PES Innovative Smart Grid Technologies (ISGT)*, pages 1–7, Jan 2012.

[40] sample residential customer load profile.

Pricing of the Geometric Asian Options Under a Multifactor Stochastic Volatility Model

Gifty Malhotra
giftymalhotradtu@gmail.com

R. Srivastava
rsrivastava@dce.ac.in

H.C. Taneja
hctaneja@dce.ac.in

Department of Applied Mathematics, Delhi Technological University,
Delhi(India)-110042

arXiv:1912.10640v1 [q-fin.PR] 23 Dec 2019

Abstract

This paper focuses on the pricing of continuous geometric Asian options (GAOs) under a multifactor stochastic volatility model. The model considers fast and slow mean reverting factors of volatility, where slow volatility factor is approximated by a quadratic arc. The asymptotic expansion of the price function is assumed, and the first order price approximation is derived using the perturbation techniques for both floating and fixed strike GAOs. Much simplified pricing formulae for the GAOs are obtained in this multifactor stochastic volatility framework. The zeroth order term in the price approximation is the modified Black-Scholes price for the GAOs. This modified price is expressed in terms of the Black-Scholes price for the GAOs. The accuracy of the approximate option pricing formulae is established, and the model parameter is also estimated by capturing the volatility smiles.

Keywords: Geometric Asian options; Modified Black-Scholes price; Multifactor stochastic volatility; Option pricing; Slow volatility factor

1 Introduction

Asian options, first introduced in 1987, are the path dependent options such that their payoff depends upon the average of the underlying asset price over the option period. Asian options are much attractive because these are generally less volatile and provide a cheaper way of hedging than their path-independent counterparts. These options are commonly traded in the currency and the commodity markets. The averaging of the underlying asset price in these options can either be arithmetic or geometric. Geometric Asian options (GAOs) are relevant and more attractive than the arithmetic ones because their closed form solutions similar to the Black-Scholes formula are available. GAOs have also been used to approximate the arithmetic Asian option prices (for instance, see [21]). Mostly the trading of the Asian options

is in a discrete sampled way but the daily sampling can be approximated by the continuous sampling case ([25]). Here the continuously sampled GAOs are considered.

Asian options have been extensively studied in literature. For the arithmetic Asian options pricing, numerical approximations are given in [21], Laplace transform formulae are derived in [10], lower and upper bounds are given in [20], Monte carlo and Laplace inversion formula is compared in [8] and a one-dimensional PDE is derived in [22]. The dimension reduction technique given in [23] is generalized in [3] by considering random volatility. The spectral expansions are discussed in [16]. The pricing based on fourier-cosine expansions is given in [11] etc. For the GAOs, explicit expressions of prices have been derived in [2, 13, 25] under the Black-Scholes framework, in [14] under the Heston stochastic volatility model, in [19] using constant elasticity of variance (CEV) model, in [24] using the perturbation techniques in a single factor stochastic model and in [12] using Barndorff-Nielsen and Shephard model under a stochastic volatility jump model framework, etc.

The constant volatility assumption of the Black-Scholes framework does not hold well in the actual market scenario. Empirically, the volatility is random which is contrary to the Black-Scholes constant volatility hypothesis. The stochastic volatility models are very well known in the option valuation as they are able to incorporate many features of volatility namely volatility smile, volatility clustering, mean reversion of volatility, etc. (see [9]). The single factor stochastic volatility models though generates the volatility smile, but are not able to explain its time varying nature. These models can be extended further to include more factors of stochastic volatility varying on the different time scales. So, the multifactor stochastic volatility models can be think of as the natural extension of these single factor stochastic volatility models. Alizadeh et al. [1] examined the dynamics of volatility and strongly recommended the two factors of volatility with one highly persistent and another quickly mean reverting factor. Extending this idea Fouque et al. [5] proposed a two factor stochastic volatility model for the pricing of European options.

Wong and Cheung [24] considered only fast mean reverting factor of volatility and using asymptotic expansions they derived the approximate pricing formula for continuous GAOs. Malhotra et al. [17] proposed a multifactor stochastic volatility model by considering both fast and slow mean reverting volatility factors, where the slow factor of volatility is approximated by a quadratic (parabolic) arc, to give the pricing formulae for the Euro-

pean options. Malhotra et al. [18] used this quadratic arc approximation of volatility for the second order moment constraint specifications to calibrate the risk-neutral density function in a two-parametric entropy framework.

In the present work, pricing formulae for the floating and fixed strike continuous GAOs are derived using a multifactor stochastic volatility model. In the GAO setup, there is the path dependence by virtue of geometric averaging of the underlying asset price. Thus, the model used for pricing is an extension of the multifactor stochastic volatility model of Malhotra et al. [17], where the pricing formulae were obtained for the path independent options. The accuracy of these formulae is also established.

The rest of the paper is organised as follows: In Section 2, multifactor stochastic volatility model to be considered for pricing is introduced. Pricing equation and the asymptotic expansion of continuous GAOs is discussed in Section 3 and 4 respectively. In Section 5, the first order approximate option pricing formulae are given. Accuracy of the approximate formulae is discussed in Section 6. The model parameter is estimated from the market by capturing the European and Asian volatility smiles together in section 7. Conclusion is given in Section 8. Detailed mathematical proofs are given in appendices.

2 Model Specification

The Asian options are the path-dependent options such that the geometric average of the underlying asset prices over the option period is required for the pricing under the GAO setup.

Let the price of an asset (non-dividend paying) at time t be X_t. The geometric average of these prices, considered in the interval $[0, t]$, is denoted by $G_{[0,t]}$ and is expressed as a stochastic process, given below.

$$G_{[0,t]} = exp\left(\frac{1}{t}\int_0^t S_\tau d\tau\right) \tag{1}$$

where $S_t = \ln X_t$. Let P^* be the risk neutral probability measure chosen by the market. Under P^*, the dynamics of stock price is described by

$$dX_t = rX_t dt + f(Y_t, Z_t)X_t dW_t^x \tag{2}$$

where r is the risk free rate of interest. Y_t and Z_t are respectively the fast and the slow mean reverting factors of stochastic volatility $f(Y_t, Z_t)$ such that

$$dY_t = \frac{1}{\epsilon}(\alpha - Y_t)dt + \frac{\nu\sqrt{2}}{\sqrt{\epsilon}}dW_t^y \tag{3}$$

and

$$Z_t = \mathcal{P}t^2 + \mathcal{Q}t + \mathcal{R} + \eta_t \tag{4}$$

In equation (3), the fast mean reverting factor Y_t follows an Ornstein-Uhlenbeck (OU) process with the rate of mean reversion $1/\epsilon$, long run mean value α and the volatility parameter $\frac{\nu\sqrt{2}}{\sqrt{\epsilon}}$. It has the long run distribution $N(\alpha, \nu^2)$ and reverts on a short time scale ϵ.
In equation (4), the approximation of the slow factor of volatility by a quadratic (parabolic) arc is considered. η_t is the error in the approximation containing the truncation terms and randomness. The value of $\mathcal{P}, \mathcal{Q}$ and $\mathcal{R}$ is specified by the process from which Z_t is approximated. The most commonly used processes to describe the volatility dynamics are the OU process and the Cox-Ingersoll-Ross (CIR) process. Here, the dynamics of Z_t is given by an OU process

$$dZ_t = k(\alpha^{'} - Z_t)dt + \beta dW_t^z \tag{5}$$

which on simplification is reduced to (4) (for details refer to [17]) giving

$$\begin{aligned} \mathcal{P} &= \tfrac{(Z_0 - \alpha^{'})}{2}k^2 \\ \mathcal{Q} &= -(Z_0 - \alpha^{'})k \\ \mathcal{R} &= Z_0 \end{aligned} \tag{6}$$

where $\alpha^{'}$ is the long run mean value about which Z_t is reverting with the rate of mean reversion k and initial value Z_0 such that $Z_0 \neq \alpha^{'}$ to ensure that $\mathcal{P} \neq 0$. η_t is the error in Z_t involving its volatility parameter β. W_t^x, W_t^y and W_t^z are the standard Brownian motions with the correlation structure:

$$E[dW_t^x.dW_t^y] = \rho_{xy}dt$$

$$E[dW_t^x.dW_t^z] = \rho_{xz}dt$$

$$E[dW_t^y.dW_t^z] = \rho_{yz}dt$$

To ensure the positive definiteness of the covariance matrix of the three Brownian motions, the correlation coefficients ρ_{xy}, ρ_{xz} and ρ_{yz} satisfy $\rho_{xy}^2 < 1, \rho_{xz}^2 < 1, \rho_{yz}^2 < 1$ and $1 + 2\rho_{xy}\rho_{xz}\rho_{yz} - \rho_{xy}^2 - \rho_{xz}^2 - \rho_{yz}^2 > 0$.

So, the above system of equations specify the multifactor stochastic volatility model to be considered for the pricing of the GAOs. The PDE for the pricing of GAOs is given in the next section.

3 PDE for the Pricing of Geometric Asian Options

Let $C^\epsilon(t, x, y, z, G)$ be the price of a continuous GAO (call) with payoff $h(X_T, G_{[0,T]})$ or $h(K, G_{[0,T]})$ (Notation C^ϵ_{fl} and C^ϵ_{fix} will be used for floating and fixed strike GAOs respectively). Using the Feynmann Kac formula ([7], pp48) , the partial differential equation governing the price comes out to be

$$\begin{aligned}\Big(&\tfrac{\partial}{\partial t} + \tfrac{1}{2}f^2(y,z)x^2\tfrac{\partial^2}{\partial x^2} + r(x\tfrac{\partial}{\partial x} - .) + \rho_{xz}\beta f(y,z)x\tfrac{\partial^2}{\partial x \partial z} + \tfrac{1}{2}\beta^2\tfrac{\partial^2}{\partial z^2}\\ &+k(\alpha^{'} - z)\tfrac{\partial}{\partial z} + \tfrac{G}{t}\ln(\tfrac{x}{G})\tfrac{\partial}{\partial G} + \tfrac{1}{\epsilon}[(\alpha - y)\tfrac{\partial}{\partial y} + \nu^2\tfrac{\partial^2}{\partial y^2}]\\ &+\tfrac{1}{\sqrt{\epsilon}}[\rho_{xy}\nu\sqrt{2}f(y,z)x\tfrac{\partial^2}{\partial x \partial y} + \rho_{yz}\nu\sqrt{2}\beta\tfrac{\partial^2}{\partial y \partial z}]\Big)C^\epsilon = 0 \qquad (7)\end{aligned}$$

with the boundary conditions

$$\begin{aligned}C^\epsilon_{fl}(T, X_T, Y_T, Z_T, G_{[0,T]}) &= h(X_T, G_{[0,T]}) = max(X_T - G_{[0,T]}, 0)\\ C^\epsilon_{fix}(T, X_T, Y_T, Z_T, G_{[0,T]}, K) &= h(K, G_{[0,T]}) = max(G_{[0,T]} - K, 0) \qquad (8)\end{aligned}$$

for the floating and fixed strike GAOs respectively. K is the strike price.

Using the transformation $s = \ln x$ and $u = t\ln\frac{G}{x}$ in (7) and after solving, we get

$$\begin{aligned}&\Big(\tfrac{\partial}{\partial t} + \tfrac{1}{2}f^2(y,z)\Big(\tfrac{\partial}{\partial s} - t\tfrac{\partial}{\partial u}\Big)^2 + \Big(r - \tfrac{1}{2}f^2(y,z)\Big)\Big(\tfrac{\partial}{\partial s} - t\tfrac{\partial}{\partial u}\Big) - r\\ &+\rho_{sz}\beta f(y,z)\Big(\tfrac{\partial}{\partial s} - t\tfrac{\partial}{\partial u}\Big)\tfrac{\partial}{\partial z} + \tfrac{1}{2}\beta^2\tfrac{\partial^2}{\partial z^2} + k(\alpha^{'} - z)\tfrac{\partial}{\partial z} + \tfrac{1}{\epsilon}[(\alpha - y)\tfrac{\partial}{\partial y} + \nu^2\tfrac{\partial^2}{\partial y^2}]\\ &+\tfrac{1}{\sqrt{\epsilon}}[\rho_{sy}\nu\sqrt{2}f(y,z)\Big(\tfrac{\partial}{\partial s} - t\tfrac{\partial}{\partial u}\Big)\tfrac{\partial}{\partial y} + \rho_{yz}\nu\sqrt{2}\beta\tfrac{\partial^2}{\partial y \partial z}]\Big)C^\epsilon(t, s, y, z, u) = 0\end{aligned}$$

It can be written as

$$\mathcal{L}^{\epsilon}C^{\epsilon}(t,s,y,z,u)=0 \tag{9}$$

where

$$\mathcal{L}^{\epsilon}=\frac{1}{\epsilon}\mathcal{L}_0+\frac{1}{\sqrt{\epsilon}}\mathcal{L}_1+\mathcal{L}_2$$

such that

$$\mathcal{L}_0=(\alpha-y)\frac{\partial}{\partial y}+\nu^2\frac{\partial^2}{\partial y^2}, \tag{10}$$

$$\mathcal{L}_1=\nu\sqrt{2}\bigg[\rho_{sy}f(y,z)\bigg(\frac{\partial}{\partial s}-t\frac{\partial}{\partial u}\bigg)\frac{\partial}{\partial y}+\rho_{yz}\beta\frac{\partial^2}{\partial y\partial z}\bigg], \tag{11}$$

and

$$\begin{aligned}\mathcal{L}_2=\ & \tfrac{\partial}{\partial t}+\tfrac{1}{2}f^2(y,z)\bigg(\tfrac{\partial}{\partial s}-t\tfrac{\partial}{\partial u}\bigg)^2+\bigg(r-\tfrac{1}{2}f^2(y,z)\bigg)\bigg(\tfrac{\partial}{\partial s}-t\tfrac{\partial}{\partial u}\bigg)-r\\ &+\rho_{sz}\beta f(y,z)\bigg(\tfrac{\partial}{\partial s}-t\tfrac{\partial}{\partial u}\bigg)\tfrac{\partial}{\partial z}+\tfrac{1}{2}\beta^2\tfrac{\partial^2}{\partial z^2}+k(\alpha'-z)\tfrac{\partial}{\partial z}.\end{aligned} \tag{12}$$

Also, the slow factor of volatility Z_t is approximated by a quadratic arc given in (4).
therefore,

$$\tfrac{\partial}{\partial z}=\bigg(\tfrac{1}{2\mathcal{P}t+\mathcal{Q}+\zeta_t}\bigg)\tfrac{\partial}{\partial t}$$

$$\tfrac{\partial^2}{\partial z^2}=\tfrac{1}{(2\mathcal{P}t+\mathcal{Q}+\zeta_t)^2}\bigg[\tfrac{\partial^2}{\partial t^2}-\bigg(\tfrac{2\mathcal{P}+\zeta_t'}{2\mathcal{P}t+\mathcal{Q}+\zeta_t}\bigg)\tfrac{\partial}{\partial t}\bigg]$$

where

$$\zeta_t=\frac{\partial\eta_t}{\partial t}$$

On substituting this value of $\frac{\partial}{\partial z}$ and $\frac{\partial^2}{\partial z^2}$ in (11) and (12), we get

$$\mathcal{L}_1=\nu\sqrt{2}\bigg[\rho_{sy}f(y,z)\bigg(\frac{\partial}{\partial s}-t\frac{\partial}{\partial u}\bigg)\frac{\partial}{\partial y}+\frac{\rho_{yz}}{2\mathcal{P}t+\mathcal{Q}+\zeta_t}\beta\frac{\partial^2}{\partial y\partial t}\bigg], \tag{13}$$

and

$$\begin{aligned}\mathcal{L}_2=&\bigg(1+\tfrac{k(\alpha'-z)}{2\mathcal{P}t+\mathcal{Q}+\zeta_t}-\tfrac{1}{2}\beta^2\tfrac{2\mathcal{P}+\zeta_t'}{(2\mathcal{P}t+\mathcal{Q}+\zeta_t)^3}\bigg)\tfrac{\partial}{\partial t}+\tfrac{1}{2}f^2(y,z)\bigg(\tfrac{\partial}{\partial s}-t\tfrac{\partial}{\partial u}\bigg)^2\\ &+\bigg(r-\tfrac{1}{2}f^2(y,z)\bigg)\bigg(\tfrac{\partial}{\partial s}-t\tfrac{\partial}{\partial u}\bigg)-r+\tfrac{\rho_{sz}\beta f(y,z)}{2\mathcal{P}t+\mathcal{Q}+\zeta_t}\bigg(\tfrac{\partial}{\partial s}-t\tfrac{\partial}{\partial u}\bigg)\tfrac{\partial}{\partial t}\\ &+\tfrac{1}{2}\beta^2\tfrac{1}{(2\mathcal{P}t+\mathcal{Q}+\zeta_t)^2}\tfrac{\partial^2}{\partial t^2}\end{aligned} \tag{14}$$

So, the required pricing equation is (9) where $\mathcal{L}_0, \mathcal{L}_1$ and $\mathcal{L}_2$ are given by (10), (13) and (14) respectively.

4 Asymptotic Price Approximation

Consider the asymptotic expansion of the Asian call option price C^ϵ in the powers of $\sqrt{\epsilon}$. i.e.

$$C^\epsilon = C_0 + \sqrt{\epsilon}C_1 + \epsilon C_2 + ... \tag{15}$$

(for i = 0,1,2,..., the notation C_i^{fl} and C_i^{fix} will be used for the floating and fixed strike continuous GAOs respectively). Put this in the pricing equation (9) to get

$$\mathcal{L}^\epsilon(C_0 + \sqrt{\epsilon}C_1 + \epsilon C_2 + ...) = 0$$

i.e.

$$(\frac{1}{\epsilon}\mathcal{L}_0 + \frac{1}{\sqrt{\epsilon}}\mathcal{L}_1 + \mathcal{L}_2)(C_0 + \sqrt{\epsilon}C_1 + \epsilon C_2 + ...) = 0 \tag{16}$$

Equating the terms of various orders equal to zero to get
Terms of order $\frac{1}{\epsilon}$:

$$\mathcal{L}_0 C_0 = 0 \tag{17}$$

Terms of order $\frac{1}{\sqrt{\epsilon}}$:

$$\mathcal{L}_0 C_1 + \mathcal{L}_1 C_0 = 0 \tag{18}$$

Terms of order 1:

$$\mathcal{L}_0 C_2 + \mathcal{L}_1 C_1 + \mathcal{L}_2 C_0 = 0 \tag{19}$$

Terms of order $\sqrt{\epsilon}$:

$$\mathcal{L}_0 C_3 + \mathcal{L}_1 C_2 + \mathcal{L}_2 C_1 = 0 \tag{20}$$

and so on. Equation (17), using the expression of $\mathcal{L}_0$ given in (10) implies that the zeroth order term C_0 in the expansion of C^ϵ is independent of the fast volatility factor y but depends on z, the slow factor of volatility.i.e.

$$C_0 = C_0(t, s, z, u) \tag{21}$$

In (18), $\mathcal{L}_1 C_0 = 0$ because C_0 is independent of y. Therefore, we are left with

$$\mathcal{L}_0 C_1 = 0$$

which gives that the first order term C_1 in the price approximation is independent of y but depends on z i.e.

$$C_1 = C_1(t, s, z, u) \tag{22}$$

For the first order price approximation, only the expressions of C_0 and C_1 are required. Firstly the expression of C_0 is obtained. Consider the equation (19), where $\mathcal{L}_1 C_1 = 0$ because C_1 is independent of y giving

$$\mathcal{L}_0 C_2 + \mathcal{L}_2 C_0 = 0 \tag{23}$$

It is a Poisson equation in C_2 with respect to y and admits a solution only if the centering condition holds:

$$E_y[\mathcal{L}_2 C_0] = 0$$

as C_0 is independent of y

$$\Rightarrow E_y[\mathcal{L}_2] C_0 = 0 \tag{24}$$

Here, $E_y[\mathcal{L}_2]$ is the average of operator $\mathcal{L}_2$ considering the invariant distribution of y. The expression of $\mathcal{L}_2$ in (14) is bit complicated. Therefore, for simplification, neglect the error term in the approximation of the slow volatility factor assuming negligible truncation error and β almost zero. Therefore, equation (14) is reduced to the simplified form:

$$\begin{aligned}\mathcal{L}_2 = &\left(1 + \tfrac{1-kt+\frac{k^2t^2}{2}}{1-kt}\right)\tfrac{\partial}{\partial t} + \tfrac{1}{2}f^2(y,z)\left(\tfrac{\partial}{\partial s} - t\tfrac{\partial}{\partial u}\right)^2 \\ &+\left(r - \tfrac{1}{2}f^2(y,z)\right)\left(\tfrac{\partial}{\partial s} - t\tfrac{\partial}{\partial u}\right) - r\end{aligned}$$

Consider,

$$\frac{1 - kt + \frac{k^2t^2}{2}}{1 - kt} = l \tag{25}$$

where, $kt \neq 1$ and $l \neq 0$. This gives

$$\mathcal{L}_2 = (1+l)\frac{\partial}{\partial t} + \frac{1}{2}f^2(y,z)\left(\frac{\partial}{\partial s} - t\frac{\partial}{\partial u}\right)^2 + \left(r - \frac{1}{2}f^2(y,z)\right)\left(\frac{\partial}{\partial s} - t\frac{\partial}{\partial u}\right) - r \tag{26}$$

It is worth noticing that

$$\mathcal{L}_2 = \mathcal{L}_{BSA} + l\frac{\partial}{\partial t}$$

where $\mathcal{L}_{BSA}$ is the standard Black-Scholes operator for the Asian options. We call $\mathcal{L}_2$ an $l-$ modified Black-scholes operator for the Asian options. Thus,

$$E_y[\mathcal{L}_2] = (1+l)\frac{\partial}{\partial t}+\frac{1}{2}\overline{\sigma}^2(z)\left(\frac{\partial}{\partial s}-t\frac{\partial}{\partial u}\right)^2+\left(r-\frac{1}{2}\overline{\sigma}^2(z)\right)\left(\frac{\partial}{\partial s}-t\frac{\partial}{\partial u}\right)-r \quad (27)$$

is the $l-$ modified Black-Scholes operator for the Asian options with the effective volatility $\overline{\sigma}(z)$. The solution of equation (24) gives the zeroth order term C_0 of the price approximation and will be called the modified Black-Scholes price for the Asian options.
After some calculation, the expression of the modified Black-Scholes price C_0 for the Asian options is obtained in terms of the Black-scholes price B_0 for the Asian options. It is given (collectively for floating and fixed strike GAOs)by

$$C_0^{fl,fix}(t,s,z,u) = \gamma^{fl,fix}(t)B_0^{fl,fix}(t,s,\overline{\sigma}(z),u)$$

or simply,

$$C_0(t,s,z,u) = \gamma(t)B_0(t,s,\overline{\sigma}(z),u) \quad (28)$$

where

$$\gamma(t) = \left[\left(\frac{2-kT}{2-kt}\right)^{\frac{2}{k}} e^{(T-t)\frac{(2-kt)(2-kT)+2}{(2-kt)(2-KT)}}\right]^M \quad (29)$$

with $kt \neq 2$ for $0 < t \leq T$ and

$$M = \left(\frac{1}{B_0}\right)\frac{\partial B_0}{\partial t} \quad (30)$$

Here, $\frac{\partial B_0}{\partial t}$ is the Black-Scholes theta for the Asian options. $\gamma(t) = 1$ at maturity to satisfy the boundary condition

$$C_0^{fl}(T,S_T,Z_T,U_T) = h(S_T,U_T)$$

or

$$C_0^{fix}(T,S_T,Z_T,U_T,K) = h(K,S_T,U_T)$$

After obtaining the expression for C_0 we intend to find C_1. For this, Consider equation (20). It is a Poisson equation in C_3 with respect to y and admits a solution only if the centering condition holds:

$$E_y[\mathcal{L}_2C_1 + \mathcal{L}_1C_2] = 0$$

$$\Rightarrow E_y[\mathcal{L}_2]C_1 + E_y[\mathcal{L}_1 C_2] = 0 \tag{31}$$

From equation (23), $\mathcal{L}_0 C_2 = -\mathcal{L}_2 C_0$

$$\Rightarrow \mathcal{L}_0 C_2 = -(\mathcal{L}_2 C_0 - E_y[\mathcal{L}_2]C_0)$$

$$\Rightarrow C_2 = -\mathcal{L}_0^{-1}(\mathcal{L}_2 - E_y[\mathcal{L}_2])C_0$$

putting it in (31) and on solving we get

$$E_y[\mathcal{L}_2]C_1 = \mathcal{G}C_0$$

where

$$\mathcal{G} = V\left[\left(\frac{\partial}{\partial s} - t\frac{\partial}{\partial u}\right)^3 - \left(\frac{\partial}{\partial s} - t\frac{\partial}{\partial u}\right)^2\right] \tag{32}$$

and

$$V = \frac{\rho_{sy}\nu}{\sqrt{2}} E_y\left[f(y,z)\frac{\partial \phi(y,z)}{\partial y}\right] \tag{33}$$

$\phi(y,z)$ is the solution of the Poisson equation

$$\mathcal{L}_0 \phi(y,z) = f^2(y,z) - \overline{\sigma}^2(z) \tag{34}$$

Now, for the floating strike geometric Asian call options,the differentials of the zeroth order term C_0^{fl} satisfies

$$\frac{\partial^{i+j} C_0^{fl}}{\partial u^i \partial s^j} = \frac{\partial^i C_0^{fl}}{\partial u^i} \tag{35}$$

where $i, j = 0, 1, 2, ...$ (see Appendix $A.1$ for the details). Therefore, for the floating strike GAO, equation (32) is reduced to

$$\mathcal{G}^{fl} = -V\left[t\frac{\partial}{\partial u} - 2t^2\frac{\partial^2}{\partial u^2} + t^3\frac{\partial^3}{\partial u^3}\right]$$

Similarly, for the fixed strike geometric Asian call option

$$\frac{\partial^{i+j} C_0^{fix}}{\partial u^i \partial s^j} = T^j \frac{\partial^{i+j} C_0^{fix}}{\partial u^{i+j}} \tag{36}$$

where $i, j = 0, 1, 2, ...$ (see Appendix $A.1$ for the details). Thus the corresponding $\mathcal{G}^{fix}$ will be

$$\mathcal{G}^{fix} = V\left[(T-t)^3\frac{\partial^3}{\partial u^3} - (T-t)^2\frac{\partial^2}{\partial u^2}\right]$$

Having the expressions of $\mathcal{G}^{fl}$ and $\mathcal{G}^{fix}$, C_1 is obtained for the floating and fixed strike GAO using Theorem (4.1) given below.

Theorem 4.1 *Let a function $A(t,s,z,u)$ satisfies*

$$E_y[\mathcal{L}_2]A(t,s,z,u) = \sum_{m=1}^{n} f_m(t)D_m(t,s,z,u)$$

with

$$A(T,S_T,Z_T,U_T) = 0$$

and $D_m(t,s,z,u)$ satisfies the PDE $E_y[\mathcal{L}_2]D(t,s,z,u) = 0$ for every $m = 1,2,...,n$, then A(t,s,z,u) will be of the form

$$A(t,s,z,u) = -\sum_{m=1}^{n}\left(\int_t^T \frac{f_m(\tau)}{1+l}d\tau\right)D_m(t,s,z,u)$$

where considering l from equation (25), $1+l = \frac{(2-kt)^2}{2(1-kt)}$. The proof of this theorem is given in Appendix $A.3$.

As the u-differentials of C_0 are interchangeable with the modified Black-Scholes operator $E_y[\mathcal{L}_2]$, so it is clear that the u-differentials of C_0^{fl} and C_0^{fix} will satisfy $E_y[\mathcal{L}_2]D(t,s,z,u) = 0$. Thus by above theorem,

$$C_1^{fl} = V\left[\left(\int_t^T \frac{\tau}{1+l}d\tau\right)\frac{\partial}{\partial u} - 2\left(\int_t^T \frac{\tau^2}{1+l}d\tau\right)\frac{\partial^2}{\partial u^2} + \left(\int_t^T \frac{\tau^3}{1+l}d\tau\right)\frac{\partial^3}{\partial u^3}\right]C_0^{fl} \tag{37}$$

and

$$C_1^{fix} = V\left[\left(\int_t^T \frac{(T-\tau)^2}{1+l}d\tau\right)\frac{\partial^2}{\partial u^2} - \left(\int_t^T \frac{(T-\tau)^3}{1+l}d\tau\right)\frac{\partial^3}{\partial u^3}\right]C_0^{fix} \tag{38}$$

$$\Rightarrow C_1^{fl} = V\left[I_1\frac{\partial}{\partial u} - 2I_2\frac{\partial^2}{\partial u^2} + I_3\frac{\partial^3}{\partial u^3}\right]C_0^{fl} \tag{39}$$

and

$$C_1^{fix} = V\left[I_4\frac{\partial^2}{\partial u^2} - I_5\frac{\partial^3}{\partial u^3}\right]C_0^{fix} \tag{40}$$

where

$$I_0 = \frac{-2}{k}\left[k(T-t)(\frac{1}{(2-kT)(2-kt)}) + \ln\frac{2-kT}{2-kt}\right]$$

$$I_1 = \frac{-2}{k^2}\left[k(T-t)(1+\frac{2}{(2-kT)(2-kt)}) + 3\ln\frac{2-kT}{2-kt}\right]$$

$$I_2 = \frac{-2}{k^3}\left[\frac{k^2}{2}(T^2-t^2) + k(T-t)(3+\frac{4}{(2-kT)(2-kt)}) + 8\ln\frac{2-kT}{2-kt}\right]$$

$$I_3 = \frac{-2}{k^4}\left[\frac{k^3}{3}(T^3-t^3) + \frac{3k^2}{2}(T^2-t^2) + 8k(T-t)(1+\frac{1}{(2-kT)(2-kt)}) + 20\ln\frac{2-kT}{2-kt}\right]$$

$$I_4 = I_2 - 2TI_1 + T^2I_0$$

$$I_5 = -I_3 + 3TI_2 - 3T^2I_1 + T^3I_0 \tag{41}$$

Clearly, C_1^{fl} and C_1^{fix} satisfy the boundary condition

$$C_1(T, S_T, Z_T, U_T) = 0$$

Collectively, for floating and fixed strike GAO

$$C_1^{fl,fix} = U^{fl,fix}C_0^{fl,fix}$$

The expression of U is given in equations (39) and (40) for the floating and fixed strike GAOs respectively.

5 First Order Approximated Price

The first order approximation to the asymptotic expansion of the GAO call price $C^{\epsilon}_{fl,fix}$ (combined for floating and fixed strike options) is denoted by $\hat{C}^{\epsilon}_{fl,fix}$ and is given by

$$C^{\epsilon}_{fl,fix} \approx \hat{C}^{\epsilon}_{fl,fix} = C_0^{fl,fix} + \sqrt{\epsilon}C_1^{fl,fix} \tag{42}$$

where $C_0^{fl,fix}$ is given by equation (28)and $C_1^{fl,fix}$ is given by equations (39) and (40) for floating and fixed strike options respectively. Unlike [24], where the approximate price was perturbed around the Black-Scholes price for the GAOs, here the first order approximated price is perturbed around the modified Black-Scholes price $C_0^{fl,fix}$ for both floating and fixed strike options. The price is modified by a modification factor $\gamma(t)$ given in (29). Also, the prices are calculated at the effective volatility $\overline{\sigma}(z)$ which is estimated by the quadratic arc and not through the standard deviation of the time series of the underlying asset returns.

6 Accuracy of the Price Approximation

The accuracy of the first order price approximation formula (42) is discussed in this section for the non-smooth pay-off of floating and fixed strike geometric Asian options. The method employed here is on the lines of [6, 4] where the accuracy of the first and the second order asymptotics is given for the European vanilla options. The similar method is also employed in [17, 24]. The non-smooth pay-off h is regularized by replacing it with modified Black-Scholes formula for the geometric Asian options with time to maturity $\Delta > 0$ and volatility $\overline{\sigma}(z)$ i.e.

$$\begin{aligned} h^{\Delta}(s,z,u) &= C_{MA}(\Delta, s, u, \overline{\sigma}(z)) \\ &= \gamma(\Delta) B_A(\Delta, s, u, \overline{\sigma}(z)) \end{aligned} \tag{43}$$

where C_{MA} is a modified Black-Scholes price for the Asian options which is represented in terms of the Black-Scholes price for Asian options B_A with modification factor $\gamma(\Delta)$ and volatility $\overline{\sigma}(z)$. Corresponding to the regularized pay-off h^{Δ}, the regularized option price $C^{\epsilon,\Delta}$ will satisfy

$$\mathcal{L}^{\epsilon} C^{\epsilon,\Delta} = 0$$

where

$$C^{\epsilon,\Delta}(T, s, y, z, u) = h^{\Delta}(s, z, u)$$

and its first order price approximation $\hat{C}^{\epsilon,\Delta}$ will be

$$\hat{C}^{\epsilon,\Delta} = C_0^{\Delta} + \sqrt{\epsilon} C_1^{\Delta}$$

with

$$C_0^{\Delta}(t, s, z, u) = C_0(t - \Delta, s, z, u) = C_{MA}(T - t + \Delta, s, u, \overline{\sigma}(z))$$

and

$$C_1^{\Delta}(t, s, z, u) = U^{\Delta} C_0^{\Delta}(t, s, z, u)$$

Theorem 6.1 *Let the volatility function $f(y, z)$ is measurable and bounded away from zero. Then for a fixed $t(< T), s, y, z, u \in \Re$, the accuracy of first order price approximation for geometric Asian option is given by*

$$|C^{\epsilon}(t, s, y, z, u) - \hat{C}^{\epsilon}(t, s, z, u)| \leq b(\epsilon^{\frac{1+g}{2}})$$

for $g < 1$ and some constant b.

The proof of this theorem is a direct implication of three lemmmas given below

Lemma 6.1.1 *Considering a fixed point* (t, s, y, z, u) *where* $t < T$*, there exist constants* $\Delta_1 > 0, \epsilon_1 > 0$ *and* $\overline{b}_1 > 0$ *so that*

$$|C^{\epsilon}(t, s, y, z, u) - C^{\epsilon,\Delta}(t, s, y, z, u)| \leq \overline{b}_1 \Delta$$

for all $0 < \Delta \leq \Delta_1$ *and* $0 < \epsilon \leq \epsilon_1$*.*

It gives that the difference in actual price C^{ϵ} and the regularized price $C^{\epsilon,\Delta}$ is small. Its proof is given in Appendix $B.1$.

Lemma 6.1.2 *Considering a fixed point* (t, s, y, z, u) *where* $t < T$*, there exist constants* $\Delta_2 > 0, \epsilon_2 > 0$ *and* $\overline{b}_2 > 0$ *so that*

$$|\hat{C}^{\epsilon}(t, s, z, u) - \hat{C}^{\epsilon,\Delta}(t, s, z, u)| \leq \overline{b}_2 \Delta$$

for all $0 < \Delta \leq \Delta_2$ *and* $0 < \epsilon \leq \epsilon_2$*.*

It gives that the difference in approximated price C^{ϵ} and its corresponding regularized price $C^{\epsilon,\Delta}$ is small. Its proof is given in Appendix $B.2$.

Lemma 6.1.3 *Considering a fixed point* (t, s, y, z, u) *where* $t < T$*, there exist constants* $\Delta_3 > 0, \epsilon_3 > 0$ *and* $\overline{b}_3 > 0$ *so that*

$$|C^{\epsilon,\Delta}(t, s, y, z, u) - \hat{C}^{\epsilon,\Delta}(t, s, z, u)| \leq \overline{b}_3 \epsilon^{\frac{1+g}{2}}$$

for all $0 < \epsilon \leq \epsilon_3$*, any* $g < 1$ *and uniformly in* $\Delta \leq \Delta_3$

It gives that the difference in regularized price $C^{\epsilon,\Delta}$ and the corresponding regularized first order price approximation $\hat{C}^{\epsilon,\Delta}$ is small. Its proof is on the lines of lemma $B.3$ of [4] and is given in Appendix $B.3$.
Using the above mentioned lemmas 6.1.1, 6.1.2 and 6.1.3 the proof of theorem 6.1 is straightforward. Consider a fixed point (t, s, y, z, u). Take $\overline{\epsilon} = \min(\epsilon_1, \epsilon_2, \epsilon_3)$ and $\overline{\Delta} = \min(\Delta_1, \Delta_2, \Delta_3)$.

$$\begin{aligned} |C^{\epsilon} - \hat{C}^{\epsilon}| &\leq |C^{\epsilon} - C^{\epsilon,\Delta}| + |C^{\epsilon,\Delta} - \hat{C}^{\epsilon,\Delta}| + |\hat{C}^{\epsilon,\Delta} - \hat{C}^{\epsilon}| \\ &\leq \overline{b}_1 \Delta + \overline{b}_2 \Delta + \overline{b}_3 (\epsilon^{\frac{1+g}{2}}) \\ &\leq 2max(\overline{b}_1, \overline{b}_2)\Delta + \overline{b}_3 (\epsilon^{\frac{1+g}{2}}) \end{aligned}$$

Let $\overline{b}_4 = max(\overline{b}_1, \overline{b}_2)$ and choose $\Delta = \epsilon$ to get

$$\begin{aligned} |C^{\epsilon} - \hat{C}^{\epsilon}| &\leq 2\overline{b}_4 \epsilon + \overline{b}_3 (\epsilon^{\frac{1+g}{2}}) \\ &\leq b(\epsilon^{\frac{1+g}{2}}) \end{aligned} \tag{44}$$

This completes the required proof.

7 Estimation of Model Parameter

After establishing the accuracy of the approximated pricing formula of GAOs, we intend to estimate V, the only parameter required for the implementation of the approximate formulae. The floating strike call and the fixed strike put options always have the positive vegas. Therefore, the volatility smiles can be captured for these options unlike the floating strike put and the fixed strike call options which may have the negative vegas. The European and Asian smiles are captured together on the lines of method given in [24]. The European smiles with the model under consideration have already been obtained in [17].

Firstly the case of floating strike call options is considered. Assume the asymptotic expansion of market implied volatility (I) for the floating strike options i.e.

$$I^{fl} = I_0^{fl} + \sqrt{\epsilon} I_1^{fl} + \epsilon I_2^{fl} + ...$$

The implied volatility I^{fl} is obtained by matching the modified Black-Scholes price for Asian options with the price C_{fl}^{ϵ} of GAO. This gives

$$C_{fl}^{\epsilon}(t, s, z, u, \overline{\sigma}(z)) = C_0^{fl}(t, s, z, u, \sigma = I^{fl})$$

Consider the Taylor expansion of the modified Black-scholes price C_0^{fl} about I_0^{fl} and the asymptotic expansion of C_{fl}^{ϵ} given in (15) to get

$$C_0^{fl}(\overline{\sigma}(z)) + \sqrt{\epsilon} C_1^{fl}(\overline{\sigma}(z))... = C_0^{fl}(I_0^{fl}) + \frac{\partial C_0^{fl}(I_0^{fl})}{\partial \sigma} \sqrt{\epsilon} I_1^{fl} + ...$$

matching the terms of order zero and ϵ on both side to get

$$I_0^{fl} = \overline{\sigma}(z)$$

and

$$I_1^{fl} = \left(\frac{\partial C_0^{fl}(I_0^{fl})}{\partial \sigma} \right)^{-1} C_1^{fl}(\overline{\sigma}(z))$$

The first order approximated value of implied volatility is

$$I^{fl} \approx I_0^{fl} + \sqrt{\epsilon} I_1^{fl}$$

For I_1^{fl}, consider the expression of C_1^{fl} from equation (39) and on rearranging the terms

$$I^{fl} \approx \overline{\sigma}(z) + \sqrt{\epsilon} \left[\left(\frac{\partial C_0^{fl}(\overline{\sigma}(z))}{\partial \sigma} \right)^{-1} \left(a_{fl} \frac{(r\overline{\sigma}(z))(I_1 \frac{\partial}{\partial u} - 2I_2 \frac{\partial^2}{\partial u^2} + I_3 \frac{\partial^3}{\partial u^3}) C_0^{fl}}{\frac{1}{(T-t)} [\frac{1}{k} \log |\frac{kT-2}{kt-2}| + \frac{T-t}{(kT-2)(kt-2)}]} + d_{fl} \right) \right] \quad (45)$$

$\Rightarrow$

$$(I^{fl}-\overline{\sigma}(z))\left(\frac{\partial C_0^{fl}(\overline{\sigma}(z))}{\partial\sigma}\right)\approx a_{fl}^{\epsilon}\left(\frac{(r\overline{\sigma}(z))(I_1\frac{\partial}{\partial u}-2I_2\frac{\partial^2}{\partial u^2}+I_3\frac{\partial^3}{\partial u^3})C_0^{fl}}{\frac{1}{(T-t)}[\frac{1}{k}\log\frac{2-kT}{2-kt}+\frac{T-t}{(2-kT)(2-kt)}]}\right)+d_{fl}^{\epsilon} \quad (46)$$

where $a_{fl}^{\epsilon}=\sqrt{\epsilon}a_{fl}$ and $d_{fl}^{\epsilon}=\sqrt{\epsilon}d_{fl}$. The expressions of $\frac{\partial C_0^{fl}}{\partial u}$, $\frac{\partial^2 C_0^{fl}}{\partial u^2}$, $\frac{\partial^3 C_0^{fl}}{\partial u^3}$ and $\frac{\partial C_0^{fl}}{\partial\sigma}$ are given in Appendix $A.2$.
a_{fl}^{ϵ} and d_{fl}^{ϵ} are estimated using the simple linear regression and the value of V is obtained from

$$V=\frac{a_{fl}(2r\sigma)}{\frac{2}{(T-t)}[\frac{1}{k}\log\frac{2-kT}{2-kt}+\frac{T-t}{(2-kT)(2-kt)}]} \quad (47)$$

Similarly, for the fixed strike put options

$$(I^{fix}-\overline{\sigma}(z))\left(\frac{\partial P_0^{fix}(\overline{\sigma}(z))}{\partial\sigma}\right)\approx a_{fix}^{\epsilon}\left(\frac{(r\overline{\sigma}(z))(I_4\frac{\partial^2}{\partial u^2}-I_5\frac{\partial^3}{\partial u^3})P_0^{fix}}{\frac{1}{(T-t)}[\frac{1}{k}\log\frac{2-kT}{2-kt}+\frac{T-t}{(2-kT)(2-kt)}]}\right)+d_{fix}^{\epsilon} \quad (48)$$

where $a_{fix}^{\epsilon}=\sqrt{\epsilon}a_{fix}$ and $d_{fix}^{\epsilon}=\sqrt{\epsilon}d_{fix}$. The expressions of $\frac{\partial^2 P_0^{fix}}{\partial u^2}$, $\frac{\partial^3 P_0^{fix}}{\partial u^3}$ and $\frac{\partial P_0^{fix}}{\partial\sigma}$ are given in Appendix $A.2$.
a_{fix}^{ϵ} and d_{fix}^{ϵ} are estimated using the simple linear regression and the value of V is obtained from

$$V=\frac{a_{fix}(2r\sigma)}{\frac{2}{(T-t)}[\frac{1}{k}\log\frac{2-kT}{2-kt}+\frac{T-t}{(2-kT)(2-kt)}]} \quad (49)$$

Numerical Illustration: The data[1] of S&P 500 index options with the floating strike is considered for the period January $4,2016$ to July $4,2016$ with the initial stock price $X_0=2013.99$. The other parameters are $k=2, r=0.0264, \epsilon=0.001, Z_0=0.1834, \alpha^{'}=0.20$. $\overline{\sigma}(z)$ is estimated from the quadratic arc $\mathcal{P}t^2+\mathcal{Q}t+\mathcal{R}$. $T-t$ ranges from 0.01 to 0.5. Firstly, V is calibrated from data. For this, consider equation (46)which can be written as

$$\mathcal{Y}\approx a_{fl}^{\epsilon}\mathcal{X}+d_{fl}^{\epsilon} \quad (50)$$

using the simple linear regression with the least square approach, values of a_{fl}^{ϵ} and d_{fl}^{ϵ} are estimated from Fig. 1. The estimated a^{ϵ} is 0.6367 and the corresponding V^{ϵ} is calculated from (47) which lies in the range -0.009 to -0.003 for different $T-t$ values. With this calibrated V , the first order

[1]Data sharing is not applicable to this article as no new data were created or analyzed in this study.

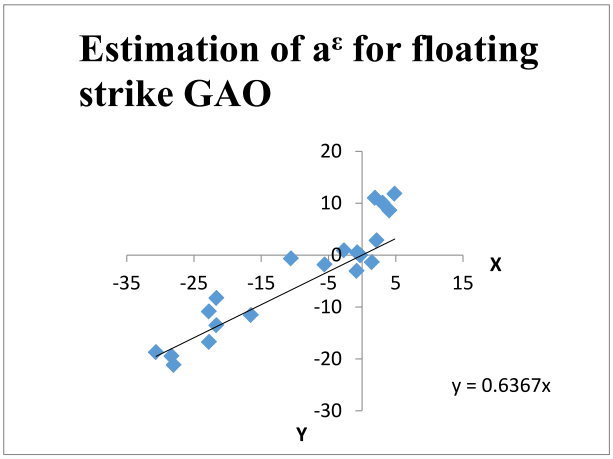


Figure 1: Estimation of a^{ϵ} from the market data of S & P 500 index implied volatility for the options with different maturity in the period January 04, 2016 to July 04, 2016.

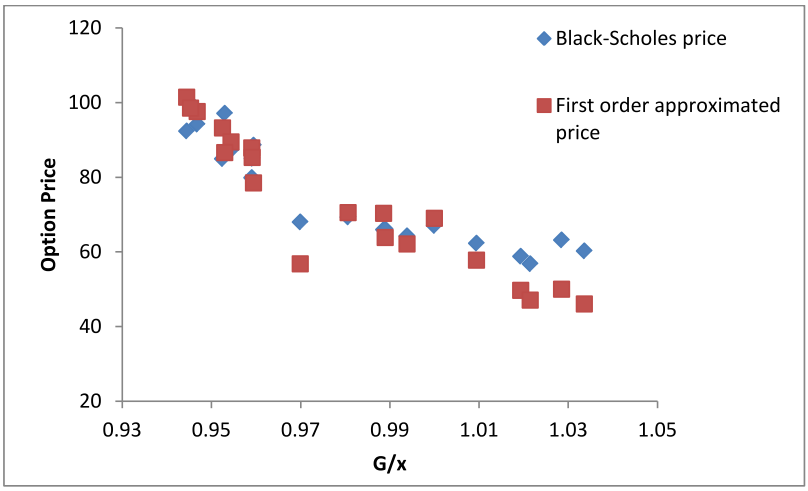


Figure 2: The comparison of Black-Scholes price of floating strike GAO with the First-order approximated price of floating strike GAO against the moneyness G/x

price approximation of the floating strike GAO is obtained from (42). The modification factor γ^{fl} is in the range 0.7 to 1.5 for the different values of $T-t$. This approximated price is compared with the Black-Scholes price for the floating strike Asian options with constant volatility 0.1834 against the

moneyness G/x shown in Fig. 2.
The effect of modification factor γ on the pricing of floating strike GAO is clearly visible in Fig. 2. For the data set under consideration, first order approximate price differs from the Black-Scholes price and the difference is more prominent for moneyness G/x greater than 1 giving an improvement over the Black-Scholes prices. The calculations get simplified with the quadratic arc approximation of the persistent volatility factor. The formulae obtained are straight forward and easy to implement. *MATLAB 2012b* and *Excel 2013* are used for numerical computations. The run time of the approximated price formulae is also very small and it takes a fraction of second to obtain the results.

8 Conclusion

A first order approximated price expression for floating and fixed strike GAO is obtained in a multifactor stochastic volatility framework where slow volatility factor is approximated by a quadratic arc. The approximated price is perturbed around the modified Black-Scholes price for the Asian options. The accuracy of the approximated formulae is established. There is only one model parameter V which is calibrated by capturing Asian and European smiles together. The first order approximate price for the floating strike GAO is obtained and compared with the corresponding Black-scholes price using the S & P 500 index Asian options data. The major advantage of using this model for pricing GAOs is that the calculations get simplified with the quadratic arc approximation of the persistent volatility factor. The pricing formulae thus obtained are straight forward and easy to implement.

Appendix A.1

The explicit formula of the modified Black-Scholes price C_0 for the floating and fixed strike geometric Asian call options is given here in terms of the Black-Scholes formula for the floating and fixed strike geometric Asian call options obtained in [15]. Firstly for the floating strike geometric Asian call, equation (28) is

$$C_0^{fl} = \gamma^{fl}(t) B_0^{fl}$$

which gives

$$C_0^{fl} = \gamma^{fl}(t)e^s\left[N(d_1) - e^{\frac{u}{T}-Q}N(d_2)\right] \tag{51}$$

where

$$\begin{aligned} d_1 &= \frac{-u+(r+\frac{\overline{\sigma}^2}{2})\frac{T^2-t^2}{2}}{\overline{\sigma}\sqrt{\frac{T^3-t^3}{3}}} \\ d_2 &= d_1 - \frac{\overline{\sigma}}{T}\sqrt{\frac{T^3-t^3}{3}} \\ Q &= (r+\tfrac{\overline{\sigma}^2}{2})\tfrac{T^2-t^2}{2T} - \tfrac{\overline{\sigma}^2}{6T^2}(T^3-t^3) \end{aligned} \tag{52}$$

with payoff function

$$h(S_T, U_T) = e^{S_T} max(1 - e^{\frac{U_T}{T}}, 0) \tag{53}$$

$\gamma^{fl}(t)$ is given in (29).

For the fixed strike geometric Asian call, equation (28) is

$$C_0^{fix} = \gamma^{fix}(t)B_0^{fix}$$

$$C_0^{fix} = \gamma^{fix}(t)\left[e^{s+\frac{u}{T}-Q}N(\hat{d}_1) - Ke^{-r(T-t)}N(\hat{d}_2)\right] \tag{54}$$

where

$$\begin{aligned} \hat{d}_2 &= \frac{\frac{u}{T}+s-\ln K+(r-\frac{\overline{\sigma}^2}{2})\frac{(T-t)^2}{2T}}{\frac{\overline{\sigma}}{T}\sqrt{\frac{(T-t)^3}{3}}} \\ \hat{d}_1 &= \hat{d}_2 + \frac{\overline{\sigma}}{T}\sqrt{\frac{(T-t)^3}{3}} \end{aligned} \tag{55}$$

with payoff function

$$h(S_T, U_T) = max(e^{S_T+\frac{U_T}{T}} - K, 0) \tag{56}$$

$\gamma^{fix}(t)$ is given in (29). It is clear from (51) that $\frac{\partial C_0^{fl}}{\partial s} = C_0^{fl}$ thus $\frac{\partial^j C_0^{fl}}{\partial s^j} = C_0^{fl}$ for every $j = 0, 1, 2, ...$ This results in (35).
Also, for the fixed strike GAOs,

$$\begin{aligned} \frac{\partial C_0^{fix}}{\partial s} &= \frac{\partial}{\partial s}(\gamma^{fix}(t)B_0^{fix}) = \gamma^{fix}(t)\frac{\partial B_0^{fix}}{\partial s} + B_0^{fix}\frac{\partial \gamma^{fix}(t)}{\partial s} \\ &= T\left(\gamma^{fix}(t)\frac{\partial B_0^{fix}}{\partial u} + B_0^{fix}\frac{\partial \gamma^{fix}(t)}{\partial u}\right) \\ &= T\frac{\partial C_0^{fix}}{\partial u} \end{aligned}$$

Thus $\frac{\partial^j C_0^{fix}}{\partial s^j} = T^j \frac{\partial^j C_0^{fix}}{\partial u^j}$. This results in (36).

Appendix A.2

For the floating strike call options: The u-differentials and the vega term is given by

$$\frac{\partial C_0^{fl}}{\partial u} = -\gamma^{fl}(t)\frac{e^{s+\frac{u}{T}-Q}}{T}N(d_2)$$

$$\frac{\partial^2 C_0^{fl}}{\partial u^2} = \frac{\gamma^{fl}(t)}{T}\left[\frac{\partial C_0^{fl}}{\partial u} + \frac{e^{s+\frac{u}{T}-Q}\Phi(d_2)}{\overline{\sigma}\sqrt{\frac{T^3-t^3}{3}}}\right]$$

$$\frac{\partial^3 C_0^{fl}}{\partial u^3} = \frac{\gamma^{fl}(t)}{T}\left[\frac{\partial^2 C_0^{fl}}{\partial u^2} + e^{s+\frac{u}{T}-Q}\left(\frac{\Phi(d_2)}{\overline{\sigma}T\sqrt{\frac{T^3-t^3}{3}}} + \frac{d_2\Phi(d_2)}{\overline{\sigma}^2\frac{T^3-t^3}{3}}\right)\right]$$

and

$$\frac{\partial C_0^{fl}}{\partial \overline{\sigma}} = \gamma^{fl}(t)e^{s+\frac{u}{T}-Q}\left[\frac{1}{T}\sqrt{\frac{T^3-t^3}{3}}\Phi(d_2) + \frac{\overline{\sigma}(T-t)^2(T+2t)N(d_2)}{6T^2}\right]$$

For the fixed strike put options: The u-differentials and the vega term is given by

$$\frac{\partial P_0^{fix}}{\partial u} = -\gamma^{fix}(t)\frac{e^{s+\frac{u}{T}-Q}}{T}N(-\hat{d}_1)$$

$$\frac{\partial^2 P_0^{fix}}{\partial u^2} = \frac{\gamma^{fix}(t)}{T}\left[\frac{\partial P_0^{fix}}{\partial u} + \frac{e^{s+\frac{u}{T}-Q}\Phi(\hat{d}_1)}{\overline{\sigma}\sqrt{\frac{(T-t)^3}{3}}}\right]$$

$$\frac{\partial^3 P_0^{fix}}{\partial u^3} = \frac{\gamma^{fix}(t)}{T}\left[\frac{\partial^2 P_0^{fix}}{\partial u^2} + e^{s+\frac{u}{T}-Q}\left(\frac{\Phi(\hat{d}_1)}{\overline{\sigma}T\sqrt{\frac{(T-t)^3}{3}}} - \frac{\hat{d}_1\Phi(\hat{d}_1)}{\overline{\sigma}^2\frac{(T-t)^3}{3}}\right)\right]$$

and

$$\frac{\partial P_0^{fix}}{\partial \overline{\sigma}} = \gamma^{fix}(t)e^{s+\frac{u}{T}-Q}\left[\frac{1}{T}\sqrt{\frac{(T-t)^3}{3}}\Phi(\hat{d}_1) + \frac{\overline{\sigma}(T-t)^2(T+2t)N(-\hat{d}_1)}{6T^2}\right]$$

where, $N(.)$ and $\Phi(.)$ are respectively the cumulative distribution function and the probability density function for the standard normal variate.

Appendix A.3

Proof of Theorem (4.1): This theorem is analogous to the theorem (3.2) of [24], which has been proved here for the modified Black-scholes operator $E_y[\mathcal{L}_2]$ which satisfies

$$E_y[\mathcal{L}_2] = [1+l]\frac{\partial}{\partial t} + \frac{1}{2}\overline{\sigma}^2(z)\left(\frac{\partial}{\partial s} - t\frac{\partial}{\partial u}\right)^2 + \left(r - \frac{1}{2}f^2(y,z)\right)\left(\frac{\partial}{\partial s} - t\frac{\partial}{\partial u}\right) - r$$

Here the case $n = 1$ is considered only and the proof will be parallel for any other value of n. Consider,

$$A(t,s,z,u) = -\left(\int_t^T \frac{f(\tau)}{1+l}d\tau\right)D(t,s,z,u)$$

Clearly,

$$A(T,s,z,u) = 0.$$

Also as

$$(1+l)\frac{\partial A}{\partial t} = f(t)D(t,s,z,u) - \left(\int_t^T \frac{f(\tau)}{1+l}d\tau\right)(1+l)\frac{\partial D}{\partial t};$$

and

$$\left(\frac{\partial}{\partial s} - t\frac{\partial}{\partial u}\right)^i A = -\left(\int_t^T \frac{f(\tau)}{1+l}d\tau\right)\left(\frac{\partial}{\partial s} - t\frac{\partial}{\partial u}\right)^i D, \forall i,$$

Therefore,

$$\begin{aligned} E_y[\mathcal{L}_2]A(t,s,z,u) &= f(t)D(t,s,z,u) - \left(\textstyle\int_t^T \frac{f(\tau)}{1+l}d\tau\right)E_y[\mathcal{L}_2]D(t,s,z,u) \\ &= f(t)D(t,s,z,u) \end{aligned}$$

because $E_y[\mathcal{L}_2]D(t,s,z,u) = 0$ This completes the proof of Theorem(4.1)

Appendix B.1

Proof of lemma (6.1.1): Proof of this lemma is on the lines of lemma 6.1.1 of [17]. Their technique is similar to [6].
As the payoff functions (53) and 56) involves both S_T and U_T, Therefore,

the joint distribution of S_T and U_T is required for the risk valuation argument. To reduce the complexity, the joint distribution of $\ln X_T$ and $\ln G_{[0,T]}$ is considered in place of the joint distribution of S_T and U_T. For the Simplification, it is considered that $\rho_{sy} = \rho$, $\rho_{yz} = 0$ and $\rho_{sz} = 0$. The proof involves the consideration of the dynamics of a new stochastic process $\ln \overline{X}_t$ which is analogous to the process $\ln X_t$ such that

$$d \ln \overline{X}_t = \left(r - \frac{1}{2} \overline{f}(t,y,z)^2 \right) dt + \overline{f}(t,y,z)(\rho dW_t^y + \sqrt{1-\rho^2} d\tilde{W}_t^x)$$

W_t^y and $\tilde{W}_t^x$ are the independent Brownian motions and

$$\overline{f}(t,y,z) = \begin{cases} f(y,z), & \text{if } t < T \\ \overline{\sigma}(z) & \text{if } t \geq T. \end{cases}$$

using the risk neutral valuation under the expectation E^*,

$$\begin{aligned} C^{\epsilon}(t,x,y,z,G) - C^{\epsilon,\Delta}(t,x,y,z,G) &= E^*_{t,x,y,z,G}[e^{-r(T-t)} h(\overline{X}_T, \overline{G}_{[0,T]})] \\ &\quad - E^*_{t,x,y,z,G}[e^{-r(T-t+\Delta)} h(\overline{X}_{T+\Delta}, \overline{G}_{[0,T+\Delta]})] \end{aligned} \tag{57}$$

$\Rightarrow$

$$\begin{aligned} &C^{\epsilon}(t,x,y,z,G) - C^{\epsilon,\Delta}(t,x,y,z,G) = \\ &\quad E^*_{t,x,y,z,G}[E^*_{t,x,y,z,G}\{e^{-r(T-t)} h(\overline{X}_T, \overline{G}_{[0,T]})|(W_p^y)_{t \leq p \leq T}\} \\ &\quad - E^*_{t,x,y,z,G}\{e^{-r(T-t+\Delta)} h(\overline{X}_{T+\Delta}, \overline{G}_{[0,T+\Delta]})|(W_p^y)_{t \leq p \leq T}\}] \end{aligned} \tag{58}$$

The joint distribution of random vector $[\ln \overline{X}_T, \ln \overline{G}_{[0,T]}]$ given the path of $(W_p^y)_{t \leq p \leq T}$ is a multivariate normal distribution with a 2-dimensional mean vector $[\mu_x, \mu_G]$ and a 2×2covariance matrix

$$\begin{bmatrix} \overline{\sigma}^2_{\rho,x}(T-t) & \overline{\sigma}^2_{\rho,x} \frac{(T-t)^2}{2T} \\ \overline{\sigma}^2_{\rho,x} \frac{(T-t)^2}{2T} & \overline{\sigma}^2_{\rho,x} \frac{(T-t)^3}{3T^2} \end{bmatrix}$$

where,

$$\mu_x = \ln \overline{X}_t + \left(r - \frac{\overline{\sigma}^2_{\rho,x}}{2} \right)(T-t) + \lambda_{t,T},$$

$$\mu_G = \frac{t}{T} \ln \overline{G}_{[0,t]} + \frac{1}{T} \left(\ln \overline{X}_t (T-t) + \left(r - \frac{\overline{\sigma}^2_{\rho,x}}{2} \right) \frac{(T-t)^2}{2} + \int_t^T \lambda_{t,q} dq \right),$$

$$\overline{\sigma}^2_{\rho,x}(T-t) = (1-\rho^2)\int_t^T \overline{f}^2(p,y,z)dp$$

and

$$\lambda_{t,T} = \rho\int_t^T \overline{f}(p,y,z)dW_p^y - \frac{\rho^2}{2}\int_t^T \overline{f}^2(p,y,z)dp$$

Similarly, The joint distribution of random vector $[\ln \overline{X}_{T+\Delta}, \ln \overline{G}_{[0,T+\Delta]}]$ given the path of $(W_p^y)_{t\le p\le T}$ is a multivariate normal distribution with a 2-dimensional mean vector $[\mu_{x,\Delta}, \mu_{G,\Delta}]$ and a 2×2covariance matrix

$$\begin{bmatrix} \overline{\sigma}^2_{\rho,x,\Delta}(T-t) & \overline{\sigma}^2_{\rho,x,G,\Delta}\frac{(T-t)^2}{2T} \\ \overline{\sigma}^2_{\rho,x,G,\Delta}\frac{(T-t)^2}{2T} & \overline{\sigma}^2_{\rho,G,\Delta}\frac{(T-t)^3}{3T^2} \end{bmatrix}$$

where,

$$\mu_{x,\Delta} = \ln \overline{X_t} + \left(r - \frac{\overline{\sigma}^2_{\rho,x,\Delta}}{2}\right)(T-t) + r\Delta + \lambda_{t,T},$$

$$\mu_{G,\Delta} = \frac{t}{T}\ln \overline{G}_{[0,t]} + \frac{1}{T}\left(\ln \overline{X}_t(T-t+\Delta) + \left(r - \frac{\overline{\sigma}^2_{\rho,x,G,\Delta}}{2}\right)\frac{(T-t)^2}{2} + r\Delta(T-t+\frac{\Delta}{2}) + \int_t^T \lambda_{t,q}dq\right),$$

with

$$\overline{\sigma}^2_{\rho,x,\Delta}(T-t) = \overline{\sigma}^2_{\rho,x}(T-t) + \overline{\sigma}^2\Delta$$

$$\overline{\sigma}^2_{\rho,x,G,\Delta}(T-t)^2 = \overline{\sigma}^2_{\rho,x}(T-t+\Delta)^2 + \rho^2\overline{\sigma}^2\Delta^2$$

and

$$\overline{\sigma}^2_{\rho,G,\Delta}(T-t)^3 = \overline{\sigma}^2_{\rho,x}(T-t+\Delta)^3 + \rho^2\overline{\sigma}^2\Delta^3$$

$\gamma(t)$ will always be finite for a finite t. Also, using the explicit Black-Scholes formula for the geometric Asian options given in Appendix $A.1$ and under the assumption of bounded $f(y,z)$, it is quite easy to show that

$$|C^\epsilon(t,x,y,z,G) - C^{\epsilon,\Delta}(t,x,y,z,G)| \le \overline{b}_1\Delta$$

for some constant $\overline{b}_1$ and some small Δ.

Appendix B.2

Proof of this lemma is on the lines of lemma 6.1.2 of [17]. Their technique is similar to [6]. The proof of the lemma is straightforward. From equations (28, 39, 40, 42), for the floating and fixed strike geometric Asian call options, the difference

$$\begin{aligned}&\hat{C}^{\epsilon}_{fl,fix}(t,s,z,u) - \hat{C}^{\epsilon,\Delta}_{fl,fix}(t,s,z,u)\\ &= \left(1 + G_1^{fl,fix}\frac{\partial}{\partial u} + G_2^{fl,fix}\frac{\partial^2}{\partial u^2} + G_3^{fl,fix}\frac{\partial^3}{\partial u^3}\right)(C_0^{fl,fix} - C_0^{fl,fix,\Delta}) \quad (59)\end{aligned}$$

where for $i = 1, 2, 3$ and $j = 0, 1, ..., 5$, $G_i^{fl,fix}$ involves I_j and V with their expressions given in (41) and (33) respectively. From [6], $V^{\epsilon} = \sqrt{\epsilon}V$ is bounded. Also for $kt \neq 2$, $\gamma(t)$, the Black-Scholes price for Asian options and its successive derivatives w.r.t u are differentiable in $t < T$, therefore the same holds for the modified Black-Scholes price for the Asian options. So it can be easily shown that

$$|\hat{C}^{\epsilon}(t,s,z,u) - \hat{C}^{\epsilon,\Delta}(t,s,z,u)| \leq \overline{b}_2\Delta$$

for some constant $\overline{b}_2$ and small Δ.

Appendix B.3

The proof of this lemma is on the lines of lemma $B.3$ of [4]. All the assumptions given in subsection (2.1) of [4] clearly holds in case of European style GAO where the volatility factors are considered to follow OU process.
Consider $C^{\epsilon,\Delta} = C_0^{\Delta} + \sqrt{\epsilon}C_1^{\Delta} + \epsilon C_2^{\Delta} + \epsilon\sqrt{\epsilon}C_3^{\Delta} - R^{\epsilon,\Delta}$
where $R^{\epsilon,\Delta}$ is the residual for the regularized problem such that

$$\mathcal{L}^{\epsilon}R^{\epsilon,\Delta} = K^{\epsilon,\Delta}$$

with

$$\begin{aligned}K^{\epsilon,\Delta} &= \mathcal{L}^{\epsilon}(C_0^{\Delta} + \sqrt{\epsilon}C_1^{\Delta} + \epsilon C_2^{\Delta} + \epsilon\sqrt{\epsilon}C_3^{\Delta} - C^{\epsilon,\Delta})\\ &= \epsilon(\mathcal{L}_2C_2^{\Delta} + \mathcal{L}_1C_3^{\Delta}) + \epsilon\sqrt{\epsilon}(\mathcal{L}_2C_3^{\Delta}) \quad (60)\end{aligned}$$

Where,

$$C_2^{\Delta} = -\frac{1}{2}\phi(y,z)D_2C_0^{\Delta}$$

$$C_3^{\Delta} = \frac{\mu}{\sqrt{2}}\rho\psi_1(y,z)D_1D_2C_0^{\Delta} - \frac{1}{2}\phi(y,z)D_2C_1^{\Delta}$$

$$D_1 = \left(\frac{\partial}{\partial s} - t\frac{\partial}{\partial u}\right)$$

$$D_2 = \left(\frac{\partial}{\partial s} - t\frac{\partial}{\partial u}\right)^2 - \left(\frac{\partial}{\partial s} - t\frac{\partial}{\partial u}\right)$$

and $\psi_1(y,z)$ is the solution of Poisson equation

$$\mathcal{L}_0\psi(y,z) = f\phi^{'} - E_y[f\phi^{'}]$$

on putting these values in (60) and on solving for the floating and fixed strike GAOs, we get

$$K_{fl}^{\epsilon,\Delta} = \gamma^{fl}\Bigg[\epsilon\Bigg(\sum_{i=1}^{6}\chi_{i,0}^{fl}(t,T,y,z)\frac{\partial^i}{\partial u^i} + (T-t)\sum_{i=1}^{6}\chi_{i,1}^{fl}(t,T,y,z)\frac{\partial^i}{\partial u^i}\Bigg)$$
$$+\epsilon\sqrt{\epsilon}\Bigg(\sum_{i=1}^{7}\chi_{i,2}^{fl}(t,T,y,z)\frac{\partial^i}{\partial u^i} + (T-t)\sum_{i=1}^{7}\chi_{i,3}^{fl}(t,T,y,z)\frac{\partial^i}{\partial u^i}\Bigg)\Bigg]B_0^{\Delta} \quad (61)$$

and

$$K_{fix}^{\epsilon,\Delta} = \gamma^{fix}\Bigg[\epsilon\Bigg(\sum_{i=1}^{6}\chi_{i,0}^{fix}(t,T,y,z)\frac{\partial^i}{\partial u^i} + (T-t)\sum_{i=1}^{2}\chi_{i,1}^{fix}(t,T,y,z)\frac{\partial^i}{\partial u^i}$$
$$+(T-t)^2\sum_{i=1}^{5}\chi_{i,2}^{fix}(t,T,y,z)\frac{\partial^i}{\partial u^i} + (T-t)^3\sum_{i=1}^{6}\chi_{i,3}^{fix}(t,T,y,z)\frac{\partial^i}{\partial u^i}\Bigg)$$
$$+\epsilon\sqrt{\epsilon}\Bigg(\sum_{i=1}^{4}\chi_{i,4}^{fl}(t,T,y,z)\frac{\partial^i}{\partial u^i} + (T-t)\sum_{i=1}^{5}\chi_{i,5}^{fl}(t,T,y,z)\frac{\partial^i}{\partial u^i}$$
$$+(T-t)^2\sum_{i=1}^{5}\chi_{i,6}^{fl}(t,T,y,z)\frac{\partial^i}{\partial u^i} + (T-t)^3\sum_{i=1}^{6}\chi_{i,7}^{fl}(t,T,y,z)\frac{\partial^i}{\partial u^i}$$
$$+(T-t)^4\sum_{i=1}^{7}\chi_{i,8}^{fl}(t,T,y,z)\frac{\partial^i}{\partial u^i} + (T-t)^5\sum_{i=1}^{7}\chi_{i,9}^{fl}(t,T,y,z)\frac{\partial^i}{\partial u^i}\Bigg)\Bigg]B_0^{\Delta} \quad (62)$$

also at maturity,

$$R^{\epsilon,\Delta}(T) = \epsilon C_2^{\Delta}(T,s,y,z,u) + \epsilon\sqrt{\epsilon}C_3^{\Delta}(T,s,y,z,u)$$
$$= Q^{\epsilon,\Delta}(s,y,z,u)(say)$$

on considering the value of C_2^{Δ} and C_3^{Δ} for both floating and fixed strike GAO at maturity, one can easily show that

$$Q_{fl,fix}^{\epsilon,\Delta}(s,y,z,u) = \gamma^{fl,fix}\Bigg[\epsilon\sum_{i=1}^{2}q_{i,0}^{fl,fix}(T,y,z)\frac{\partial^i}{\partial u^i} + \epsilon\sqrt{\epsilon}\sum_{i=1}^{3}q_{i,1}^{fl,fix}(T,y,z)\frac{\partial^i}{\partial u^i}\Bigg]B_0^{\Delta}$$

Clearly for fixed strike options this value is zero. Using this terminal condition, residual $R^{\epsilon,\Delta}$ has the probabilistic representation

$$R^{\epsilon,\Delta} = E^*_{t,s,y,z,u}\left[e^{-r(T-t)}Q^{\epsilon,\Delta} - \int_t^T e^{-r(\tau-t)}K^{\epsilon,\Delta}d\tau\right]$$

Using lemma $B.4$ of [4], for both floating and fixed strike options at a fixed $t < T$,

$$\left|E^*_{t,s,y,z,u}\left[Q^{\epsilon,\Delta}(S_T, Y_T, Z_T, U_T)\right]\right| \leq b_3\epsilon^{\frac{1+g}{2}}$$

and

$$\left|E^*_{t,s,y,z,u}\left[\int_t^T e^{-r(\tau-t)}K^{\epsilon,\Delta}(s,y,z,u)d\tau\right]\right| \leq c_3\epsilon^{\frac{1+g}{2}}$$

therefore,

$$|C^{\epsilon,\Delta}(t,s,y,z,u) - \hat{C}^{\epsilon,\Delta}(t,s,z,u)| = |\epsilon C_2^\Delta + \epsilon\sqrt{\epsilon}C_3^\Delta - R^{\epsilon,\Delta}|$$

where C_2^Δ and C_3^Δ are bounded for $t < T$ giving

$$|C^{\epsilon,\Delta}(t,s,y,z,u) - \hat{C}^{\epsilon,\Delta}(t,s,z,u)| \leq \overline{b}_3\epsilon^{\frac{1+g}{2}}$$

for some $\overline{b}_3$ and $g < 1$.

References

[1] Alizadeh,S., Brandt,M.W., Diebold,F.X.: Range-Based Estimation of Stochastic Volatility Models. J. Finance. **57**, 1047 – 1091 (2002)

[2] Angus,J.E.: A Note on Pricing Asain Derivatives with Continuous Geometric Averaging. J. Futures Markets. **19**, 845 – 858 (1999)

[3] Fouque,J.P., Han,C.H.: Pricing Asian Options with Stochastic Volatility. Quant. Finance. **3**, 353 – 362 (2003)

[4] Fouque,J.P., Lorig,M., Sircar,R.: Second Order Multiscale Stocahstic Volatility Asymptotics: Stochastic Terminal Layer Analysis and Calibration. Finance Stoch. **20**, 543 – 588 (2016)

[5] Fouque,J.P., Papanicolau,G., Sircar,R., Solna,K.: Multiscale Stochastic Volatility Asymptotics. SIAM J. Multiscale Model. Simul. **2**, 22 – 42 (2003)

[6] Fouque,J.P., Papanicolaou,G., Sircar,R., Solna,K.: Singular Perturbations in Option Pricing. SIAM J. Appl. Mathematics **63**, 1648 – 1665 (2003)

[7] Fouque,J.P., Papanicolau,G., Sircar,R., Solna,K.: Multiscale Stochastic Volatility for Equity, Interest Rate and Credit Derivatives. Cambridge Univ. Press, Cambridge (2011)

[8] Fu,M.C., Madan,D.B., Wang,T.: Pricing Continuous Asian Options: A Comparison of Monte Carlo and Laplace Transform Inversion Methods. J. Comput. Finance. **2**, 49-74 (1999)

[9] Gatheral,J.: The Volatility Surface: A Practitioner's Guide. John Wiley and Sons,Inc., New York (2006)

[10] Geman, H., Yor, M.: Bessel Processes, Asian Options and Perpetuities. Math. Finance. **3**, 349-375 (1993)

[11] Huang,C.S., OHara,J.G., Mataramvura,S.: Efficient pricing of discrete arithmetic Asian options under mean reversion and jumps based on Fourier-cosine expansions. J. Comput. Appl. Mathematics **311**, 230 – 238 (2017)

[12] Hubalek,F., Sgarra,C.: On the Explicit Evaluation of the Geometric Asian Options in Stochastic Volatility Models with Jumps. J. Comput. Appl. Mathematics **235**, 3355 – 3365 (2011)

[13] Kemna,A.G.Z., Vorst,A.C.F.: A Pricing method for Options based on Average Asset Values. J. Bank. Finance. **14**, 113 – 129 (1990)

[14] Kim,B., Wee,I.S.: Pricing of Geometric Asian Options under Heston's Stochastic Volatility Model. Quant. Finance. **14**, 1795 – 1809 (2014)

[15] Kwok,Y.K.: Mathematical Models of Financial Derivatives (2nd edition). Springer Berlin Heidelberg (2008)

[16] Linetsky,V.: Spectral Expansions for Asian (Average Price) Options. Oper. Res. **52**, 856 – 867 (2004)

[17] Malhotra,G., Srivastava,R., Taneja,H.C.: Quadratic approximation of the slow factor of volatility in a multifactor stochastic volatility model. J Futures Markets. **38**, 607 - 624 (2018)

[18] Malhotra,G., Srivastava,R., Taneja,H.C.: Calibration of the risk-neutral density function by maximization of a two-parameter entropy. Physica A. **513**, 45 - 54 (2019)

[19] Peng,B.: Pricing Geometric Asian Options under CEV Process. Int. Econom. J. **20**, 515 – 522 (2006)

[20] Rogers, L.C.G., Shi, Z.: The value of an Asian option. J. Appl. Probab. **32**, 1077-1088 (1995)

[21] Turnbull, S.M., Wakeman, L.M.: A Quick algorithm for Pricing European Average Options. J. Financ. Quant. Anal. **26**, 377-389 (1991)

[22] Vecer,J.: A New PDE Approach for Pricing Arithmetic Average Asian Options. J. Comput. Finance. **4**, 105 – 113 (2001)

[23] Vecer,J.: Unified pricing of Asian options . Risk. 113 – 116 (2002)

[24] Wong,H.Y., Cheung,Y.L.: Geometric Asian Options: Valuation and Calibration with Stochastic Volatility. J. Financ. Econom. **4**, 301 – 314 (2004)

[25] Zhang,J.E.: Pricing Continuously Sampled Asian Options with Perturbation Method. J. Futures Markets. **23**, 535 – 560 (2003)

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/326008135

Chance Constraints Integrated MPC Navigation in Uncertainty amongst Dynamic Obstacles: An overlap of Gaussians approach

Preprint · June 2018

CITATIONS
0

READS
28

4 authors, including:



Bharath Gopalakrishnan
International Institute of Information Technology, Hyderabad
12 PUBLICATIONS **39** CITATIONS
SEE PROFILE



Madhava Krishna
International Institute of Information Technology, Hyderabad
222 PUBLICATIONS **1,162** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

MS research project on intelligent UI based navigation View project

Compliant Modular Robot View project

All content following this page was uploaded by Bharath Gopalakrishnan on 11 November 2019.

The user has requested enhancement of the downloaded file.

Probabilistic obstacle avoidance and object following: An overlap of Gaussians approach

Dhaivat Bhatt[1*] Akash Garg[2*] Bharath Gopalakrishnan[1] K. Madhava Krishna[1]

***Abstract*— Autonomous navigation and obstacle avoidance are core capabilities that enable robots to execute tasks in the real world. We propose a new approach to collision avoidance that accounts for uncertainty in the states of the agent and the obstacles. We first demonstrate that measures of entropy—used in current approaches for uncertainty-aware obstacle avoidance—are an inappropriate design choice. We then propose an algorithm that solves an *optimal* control sequence with a guaranteed risk bound, using a measure of overlap between the two distributions that represent the state of the robot and the obstacle, respectively. Furthermore, we provide closed form expressions that can characterize the overlap as a function of the control input. The proposed approach enables model-predictive control framework to generate bounded-confidence control commands. An extensive set of simulations have been conducted in various constrained environments in order to demonstrate the efficacy of the proposed approach over the prior art. We demonstrate the usefulness of the proposed scheme under tight spaces where computing risk-sensitive control maneuvers is vital. We also show how this framework generalizes to other problems, such as object-following.**

I. INTRODUCTION

Autonomous robots are required to navigate between a set of defined locations while avoiding any other agent present on its predicted path. However, uncertainty arising in the motion of a robot coupled with uncertainty in observation as well as in environment makes path planning an interesting job. This problem is necessary to address when autonomous robots and humans exist in a shared environment like semi-automated warehouses or mining sites. Strong requirement of uncertainty aware motion planning necessitates need of new algorithms that can accurately model various uncertainties arising in stochastic world of humans. Towards this, we make an effort to develop a theoretical approach that could take quantitative measure of uncertainty of robot and other objects into consideration before making a planning decision. The environment can be populated with static and dynamic objects as well as other robots. A great amount of work done in the past deals with motion planning of the robot from given start pose to goal pose in the presence of obstacles. Many of the deterministic obstacle avoidance algorithms studied in literature, when used for a robot present in an uncertain environment can lead to substantial degradation in the desired result and can even make the source robot to collide into the obstacle, in the worst case. Chance constraint is one of the efficient way of employing the probability of constraint violation which is below a user specified bound. These constraints are combined with other constraints accounting for the kinematic and dynamic feasibility in form of actuation constraints of the robot. Our approach generates smooth trajectories which can be employed on real robots. This paper has several novel findings and contributes in following ways,

- This is first such formulation, conditioned on agent and obstacle uncertainty into an MPC framework through theory of overlapping Gaussians.

[1]Affiliated with KCIS, Robotics Research Center, IIIT Hyderabad
[2]Affiliated with Delhi Technological University
*Equal contribution
This work is supported by grants made available from Rockwell Collins, India development center.

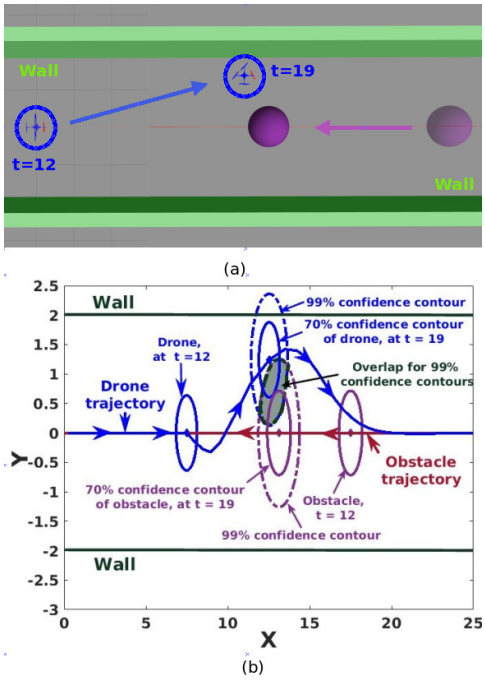


Fig. 1. We show result of our probabilistic obstacle avoidance algorithm in constrained corridor when an obstacle is approaching in antipodal configuration. Figure 1(a) shows gazebo snapshot of agent positions for two different time instances. Our objective is to ensure that 70% confidence contour of agent's Gaussian is avoiding atleast 70% confidence contour of obstacle's Gaussian distribution at any time instance during the entire trajectory. In figure 1(b), we show how the control maneuvers found through our algorithm respects user-specified safety bound. For example, at time $t = 19$, we can see clear overlap(shaded area) between 99% confidence contours of the agent and obstacle. However, confidence contours corresponding to 70% are not being penetrated which can be seen from figure 1(b), thus, we can clearly see how a user-specified lower bound is being respected.

- We demonstrate why our modeling is more consistent and appropriate compared to the widely used Bhattacharya distance, that measures the similarity of two probability distributions.
- We formulate an optimization framework that solves for chance constraints using the theory of overlapping Gaussians.
- We show effective results in various simulation settings that showcase versatility of the method. Specifically we show where the distributions are non isotropic, which is closer to real setting.

Above novelties are arising out of two important problems we are solving, First one is probabilistic collision avoidance and second is probabilistic object following. We formally define them below,

- **Probabilistic collision avoidance:**

  Having knowledge about agent's initial state, agent's belief propagation and the trajectory of obstacles' state, the objective is to find an optimal sequence of control inputs. Such a sequence ensures that overlapping area between

any obstacle and agent's probability distributions at any time instance is less than prescribed maximum allowed overlap (Υ_{max}).

- **Probabilistic object following:**

  Having knowledge about agent's initial state, agent's belief propagation and the trajectory of object's state, the objective is to find an optimal sequence of control inputs. Such a sequence ensures that overlapping area between agent and object's probability distributions at any time instance is between prescribed range of overlap. This range is characterized by lower and upper bounds of overlap $\Upsilon_{min}, \Upsilon_{max}$ respectively.

In both problem statements, state is characterized by a Gaussian distribution. Optimality is defined as a goal oriented, minimal jerk trajectory planning. Overlap between two Gaussian distributions is uniquely related to confidence contours at which they are touching each-other. Providing confidence contour is more intuitive and safe than providing raw thresholds of prescribed area of overlap. Hence, input to our algorithm would be Gaussian confidence contours which are not allowed to be penetrated, which is a user specified number.

II. Related work

This section reviews recent advances in MPC for autonomous navigation. The evident advantage of using MPC in motion planning and autonomous navigation has been well demonstrated in ([1],[2], [3], [4]) among many. Formulations along the lines of [1], [2] do a great job in achieving performance in terms of quality of trajectory, computation time and novelty of approach. However, they have been developed for a deterministic setting and hence do not take into consideration the uncertainty in state of the robot and obstacle into their collision avoidance routine. While MPC formulations along the lines of [4] does take into consideration the state uncertainty and demonstrate interesting maneuvers in complex driving scenarios, [4] considers uncertainty only in the state of the obstacle, and the collision avoidance is modeled through a Minkowski sum approach. Considering agent's uncertainty into Minkowski sum formulation would be very cumbersome as Minkowski sum between two ellipses is very complex. [5] takes into consideration the uncertainty of agent and assumes the obstacle to be static and deterministic, it models collision as a measure of entropic distance(Similar to Mahalanobis distance. It is shown in the later section of this paper that formulating probabilistic collision avoidance as an entropic distance may not be the most appropriate approach when uncertainty in the agent and obstacle is considered. [6] attempts to solve for collision avoidance in a multiagent scenario under uncertainty. It achieves it through an RRT framework and a sampling strategy to choose a path corresponding to a desired level of safety confidence. Authors in [7] solve probabilistic multi-agent motion planning problem by considering Pianka's measure([8]) to estimate overlapping probability. This measurement is similar to entropic distance which provides measure of similarity for the given Gaussian statistics. Formulating collision avoidance as a chance constraint is well explored in [9],[10] among the many. [10] demonstrates an efficient way of solving an intractable chance constraint through a series of reformulations. These were built on time scale velocity obstacle concepts [11].

There has been considerable body of work([12], [13], [14]) exploring obstacle avoidance routine through linear chance constraints. However, these methods are modeling linear chance constraints for a static and deterministic obstacles. In [12], convex polyhedral obstacles are encoded by combination of a set of feasible half-spaces, which is algebraically represented by a combination of linear constraints. Obstacle avoidance problem was solved using disjunctive linear programming. Here, risk is uniformly distributed to decompose joint chance constraint. Authors in [12] continue their work in ([13]) where they proposed a risk allocation strategy to optimally distribute risk bounds to individual chance constraints. They additionally prove convexity of iterative risk allocation method for linear systems in [14]. We provide a supplementary document where we illustrate why obstacle avoidance becomes difficult problem to tackle using linear chance constraints based framework when uncertainty is involved in obstacle and agent's states. This method will fail to provide a probabilistic safety bound due to intractability arising in collision avoidance constraint because of obstacle's uncertainty. We provide a more comprehensive explanation of this in our supplementary document. We also provide comparison between our method with bounding box approach and experimentally demonstrate superiority of our method, illustrated through diagrams. We suggest reader to go through the additional document after reading the paper. It can be found at **https://robotics.iiit.ac.in/people/dhaivat.bhatt/Supplementary.pdf**

At this point, it is important to note that we do not approximate the distribution of chance constraint as a Gaussian through a first order Taylor series approximation method, instead we come up with an alternate way of representing it as an overlap of Gaussians. This way we are differing from most of the existing approaches. We will conclude in the subsequent sections that an overlap of Gaussians method gives rise to maneuvers pertaining to user specified confidence bounds(confidence of safety). Our approach is applicable to static as well as dynamic obstacles. Our mathematical framework adapts well for 2D and 3D navigation tasks. With overlap of Gaussians based modeling, we easily avoid complexity of considering Minkowski sum between two ellipses and other approaches that model probabilistic collision avoidance through entropic distances. We further show its unique application to object following with a simple extension of collision avoidance framework. Such formulation can be used for human-following robots where there is uncertainty involved in human pose/motion estimation apart from error in agent's state estimation. To the best of author's knowledge, this is a first such attempt where non-linear chance constraints are tackled by representing them using theory of overlapping Gaussians and not any other entropic distance.

III. Prerequisite

This section describes deterministic MPC framework, formulated along the lines of [1]. A linear motion model is being considered, and collision avoidance is added as affine constraints. Entire proposed framework is solved as a sequential convex programming(SCP) routine[15].

A. Trajectory optimization in a deterministic setting

In a deterministic trajectory optimization setting, our objective is to reach desired goal in a given amount of time while ensuring a collision-free trajectory. This problem can be modeled by considering a set of cost functions and constraints.

Let the start position of an agent be $X_0 = (x_0, y_0, z_0)$. Our objective is to reach the goal position $G_f = (G_f^x, G_f^y, G_f^z)$ in N time-steps, each time-step of duration τ. Position of an agent at any time instant t_i is $X_{t_i} = (x_{t_i}, y_{t_i}, z_{t_i})$, velocity of an agent at time instant t_i is $V_{t_i} = (v_{t_i}^x, v_{t_i}^y, v_{t_i}^z)$. We have P obstacles in the environment. Their position at time t_i is

defined as $O^j_{t_i} = (o^{xj}_{t_i}, o^{yj}_{t_i}, o^{zj}_{t_i})$, for $\forall$ j= $\{1, 2, 3, ... P\}$. For static obstacles, the obstacle locations will be independent of t_i. The agent and obstacles are approximated as circular objects with radius of the agent being R_{agent} and radius of obstacle j is R_j, $\forall\ j \in \{1, 2, 3, ... P\}$.

$$\underset{V_{t_i}}{\text{argmin}} \quad J = J_{terminal} + J_{smooth} \tag{1a}$$

$$X_{t_{i+1}} = f(X_{t_i}, V_{t_i}) \tag{1b}$$

$$V_{min} \leq V_{t_i} \leq V_{max} \tag{1c}$$

$$a_{min} \leq \frac{V_{t_{i+1}} - V_{t_i}}{\tau} \leq a_{max} \tag{1d}$$

$$C^{t_i}{}_{obs_j}(x_{t_i}, y_{t_i}, z_{t_i}, o^{xj}_{t_i}, o^{yj}_{t_i}, o^{zj}_{t_i}, R_{agent}, R_j) \leq 0 \tag{1e}$$
$$\forall\ i \in \{1, 2, 3, ... N\}\ , \forall\ j \in \{1, 2, 3, ... P\}$$

The above set of equations defines the cost function as well as constraints. Equation 1a describes the objective function.

$$J_{terminal} = (x_{t_N} - G_x)^2 + (y_{t_N} - G_y)^2 + (z_{t_N} - G_z)^2 \tag{2}$$

The terminal cost forces our system to achieve goal-state(G_f) at the end of the trajectory.

$$J_{smooth} = \sum_{i\,=\,2}^{N\,-\,1} (\frac{(V_{t_{i+1}} + V_{t_{i-1}} - 2V_{t_i})}{\tau^2})^2 \tag{3}$$

The smoothness cost as described above ensures smooth trajectory with minimal jerk. It minimizes the jerk which is modeled as second order finite difference between subsequent linear velocities. This term penalizes sudden deviations in the acceleration profile and ensures smooth velocity transitions. Equation 1b is the process model of the agent. It ensures that control variables and states are adhering the motion model of the agent. Process model of holonomic agent can be described as below,

$$X_{t_i} = f(X_0, V_{t_1}, V_{t_2}, ...V_{t_i}, \tau) = X_0 + \sum_{k=1}^{i} V_{t_k}\tau \tag{4}$$

Equations 1c-1d, represents constraints modeling actuation limitations of the agent. The bounds on linear acceleration and velocity ensures that the actuation capabilities of the agent are not violated. Equation 1e models collision avoidance constraint between obstacle j and agent. For a deterministic setting, this can be a simple euclidean distance constraint as below.

$$\begin{aligned} C^{t_i}{}_{obs_j}(.) = &-(x_{t_i} - o^{xj}_{t_i})^2 - (y_{t_i} - o^{yj}_{t_i})^2 \\ &-(z_{t_i} - o^{zj}_{t_i})^2 + (R_j + R_{agent})^2 \leq 0. \\ &\forall\ i \in \{1, 2, 3, ... N\}\ , \forall\ j \in \{1, 2, 3, ... P\} \end{aligned} \tag{5}$$

Above constraint is purely non-linear in agent's position i.e. X_{t_i} = (x_{t_i}, y_{t_i}, z_{t_i}), which is our variable of interest. We linearize it along the lines of [1] and solve the proposed routine using sequential convex programming. The trajectory optimization routine is then integrated into a model predictive control framework.

IV. Collision avoidance under uncertainty

Robot motions are generally erroneous in nature. There is always some uncertainty associated with the location of the agent. Sensing modality also gives inaccurate estimate of the state of the obstacle/object. Uncertain position estimate of the obstacle leads to erroneous trajectory estimation. In such cases, the constraint in equation 1e takes the form of 6

$$\Pr(C^{t_i}{}_{obs_j}(x_{t_i}, y_{t_i}, z_{t_i}, o^{xj}_{t_i}, o^{yj}_{t_i}, o^{zj}_{t_i}, R_{agent}, R_j) \leq 0) \geq \Delta \tag{6}$$
$$\forall\ i \in \{1, 2, 3, ... N\}\ , \forall\ j \in \{1, 2, 3, ... P\}$$

Constraints of the form 6, are generally known as chance constraints, and in most cases may not have a distribution that can be computed in closed form. The nature of these chance constraints also depends on the form of the deterministic constraints that they are built on. For example chance constraint 6 arising out of 5 can be expressed as an entropic distance. Following three sections provides a comprehensive overview of three important ways to formulate a chance constraint.

A. Theoretical characterization of chance constraint

The chance constraint in 6 can take the form of a transformed distribution of 5 as shown in Equation 7,

$$\int \cdots \int_{V_j} \Pr(X_{t_i}, O^j_{t_i}) dX_{t_i} dO^j_{t_i} \tag{7}$$

Where, X_{t_i} = ($x_{t_i}, y_{t_i}, z_{t_i}$), position of the agent at time t_i. Under state uncertainty, let $X_{t_i} \sim \mathcal{N}\ (\hat{X}_{t_i}, \Sigma^d_{t_i})$ and $O^j_{t_i} \sim \mathcal{N}\ (\hat{O}^j_{t_i}, \Sigma^{o_j}_{t_i})$ be the Gaussian parameterization of the agent and obstacle j positions at time t_i. Then, $\Pr(X_{t_i}, O^j_{t_i})$ takes the following form,

$$\Pr(X_{t_i}, O^j_{t_i}) \sim \mathcal{N}(\begin{pmatrix} \hat{X}_{t_i} \\ \hat{O}^j_{t_i} \end{pmatrix}, \begin{pmatrix} \Sigma^d_{t_i} & 0 \\ 0 & \Sigma^{o_j}_{t_i} \end{pmatrix}) \tag{8}$$

Here, $\Pr(.)$ symbolizes probability of a random variable. When we substitute equation 8 in equation 7, equation 7 becomes analytically intractable. The closed form solution of equation 7 doesn't exist. Authors in [6] attempted to tackle problem of multi-robot motion planning for differential drive robots, where they numerically evaluate equation 7 over the region of interest. The region of interest here would be set of positions of the agent and obstacles for which collision occurs. Our objective would be to minimize the value of equation 7, which means we want to maximize the probability of collision avoidance. However, one drawback of this procedure is that characterization of such a region(V_j) is generally tough.

There has been a lot of work to characterize the entropic distance between two distributions. One of the commonly used techniques for entropic distances are chi-square distances, Bhattacharyya distances among the many. We describe case of Bhattacharyya distances which is an extension to Mahalanobis distance.

B. Bhattacharyya distance

Bhattacharyya distance [16] gives measure of similarity between two continuous/discrete probability distributions. It attempts to quantify the overlap between two distributions. For Gaussian distributions, Bhattacharyya distance has well-defined analytical formula to capture the overlap. For two multivariate normal distributions, p_i = $\mathcal{N}(\mu_i, \Sigma_i)$ with $i = \{1, 2\}$, Bhattacharyya distance metric is defined as below,

$$\begin{aligned} \mathcal{BC}(p_1, p_2) = &\frac{1}{8}(\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2) + \\ &\frac{1}{2} \ln \frac{det(\Sigma)}{\sqrt{det(\Sigma_1) det(\Sigma_2)}} \end{aligned} \tag{9}$$

Where, $\Sigma = (\Sigma_1 + \Sigma_2)/2$. This is an extension of Mahalanobis distance. However, equation 6 can not be completely modeled through similarity/dissimilarity given by such entropic measure. Because the notion of probability is not complete in Bhattacharyya metric and particular distance doesn't have a unique mapping to certain value in probability space. In figure 2, we demonstrate that for different Gaussian distributions with same overlap have different Bhattacharyya distances. Due to this limitation, entropic distance can't be used to model our non-linear chance constraint.

C. Theory of overlapping of Gaussians

The theory of overlap between two Gaussians has been widely studied in [17], which is built upon [18]. The authors in [18] attempts to get an optimal linear separator which minimizes misclassification error when the objective is to classify the sample as coming from one of the several populations. We briefly state the theory to get approximate estimate of component of overlap between two Gaussians.

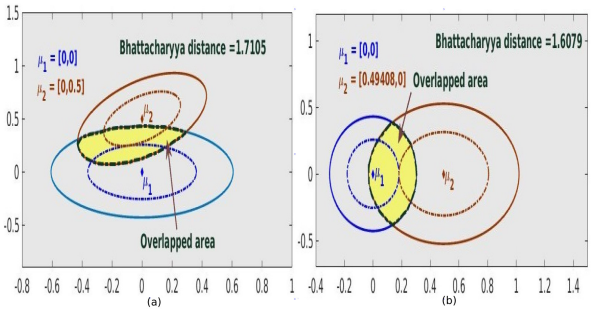


Fig. 2. Diagrammatic explanation of why Bhattacharyya distance is not appropriate metric to model chance constraint problem. In this figure, we have taken 2 sets of Gaussian distribution pairs which are touching at same confidence contour of 80.51%. Meaning, area of overlap between two Gaussians shaded in yellow is same for both sets of Gaussian distributions. For figure 2(a), $\Sigma_1 = \begin{pmatrix} 0.04 & 0 \\ 0 & 0.02 \end{pmatrix}$, $\Sigma_2 = \begin{pmatrix} 0.02 & 0.01 \\ 0.01 & 0.02 \end{pmatrix}$. Bhattacharyya distance evaluated using equation 9 for this set of co-variances turns out to be 1.7105. While, for figure 2(b), $\Sigma_1 = \begin{pmatrix} 0.01 & 0 \\ 0 & .0.02 \end{pmatrix}$, $\Sigma_2 = \begin{pmatrix} 0.03 & 0 \\ 0 & 0.03 \end{pmatrix}$. Value of Bhattacharyya distance for figure 2(b) is 1.6079. Hence, for same amount of overlap between two sets of Gaussian distributions, Bhattacharyya distances are turning out to be different. Areas shaded with yellow in figure indicate amount of overlap.

The linear separator proposed in[18] works for two Gaussians of dimension $d \geq 1$.

Let the linear separator(a hyperplane in d dimensional space) be $\alpha^T x = \beta$ where $\alpha, x \in \mathbb{R}^d$ and $\beta \in \mathbb{R}$. $\alpha^T x \leq \beta$ classifies x into first cluster and $\alpha^T x > \beta$ classifies x into second cluster. We will briefly explain the procedure to obtain α, β and estimate the area of overlap(Υ) between two Gaussian distributions $\mathcal{N}(\mu_i, \Sigma_i)$ with $i = \{1, 2\}$. Here, x is coming from one of the above two Gaussian distributions and $\alpha^T x$ is a linear transformation which transforms the original distribution into uni-variate normal distribution. The probability of misclassification when x is coming from first distribution is,

$$\mathbb{P}_1(\alpha^T x > \beta) = \mathbb{P}_1\left(\frac{\alpha^T x - \alpha^T \mu_1}{\sqrt{\alpha^T \Sigma_1 \alpha}} > \frac{\beta - \alpha^T \mu_1}{\sqrt{\alpha^T \Sigma_1 \alpha}}\right)$$
$$\mathbb{P}_1(\alpha^T x > \beta) = 1 - \Phi\left(\frac{\beta - \alpha^T \mu_1}{\sqrt{\alpha^T \Sigma_1 \alpha}}\right) = 1 - \Phi(\eta_1) = \mathbb{P}_1(\eta_1) \tag{10}$$

Similarly, probability of misclassification when sample x belongs to population 2 equals,

$$\mathbb{P}_2(\alpha^T x \leq \beta) = \mathbb{P}_2\left(\frac{\alpha^T x - \alpha^T \mu_2}{\sqrt{\alpha^T \Sigma_2 \alpha}} \leq \frac{\beta - \alpha^T \mu_2}{\sqrt{\alpha^T \Sigma_2 \alpha}}\right)$$
$$\mathbb{P}_2(\alpha^T x \leq \beta) = 1 - \Phi\left(\frac{\alpha^T \mu_2 - \beta}{\sqrt{\alpha^T \Sigma_2 \alpha}}\right) = 1 - \Phi(\eta_2) = \mathbb{P}_2(\eta_2) \tag{11}$$

Where, $\eta_1 = \frac{\beta - \alpha^T \mu_1}{\sqrt{\alpha^T \Sigma_1 \alpha}}$ and $\eta_2 = \frac{\alpha^T \mu_2 - \beta}{\sqrt{\alpha^T \Sigma_2 \alpha}}$ are two scalars. Φ in equations 10-11 denotes a cumulative distribution function for a univariate standard normal distribution. Φ is a monotonically increasing function. Our objective is following,

$$max(\mathbb{P}_1(\eta_1), \mathbb{P}_2(\eta_2)) \rightarrow \min_{\alpha \in \mathbb{R}^d \beta \in \mathbb{R}} \tag{12a}$$
$$min(\eta_1, \eta_2) \rightarrow \max_{\alpha \in \mathbb{R}^d \beta \in \mathbb{R}} \tag{12b}$$

Equations 12a-12b are equivalent due to monotonic nature of the Φ. The objective is to maximize the minimum of (η_1, η_2). The objective is to find α, β, which will minimize the maximum probability of misclassification. Analytical characterization of α, β in terms of $\mu_1, \Sigma_1, \mu_2, \Sigma_2$ can be expressed as following,

$$\alpha = (\lambda_1 \Sigma_1 + \lambda_2 \Sigma_2)^{-1}(\mu_2 - \mu_1) \tag{13a}$$
$$\beta = \alpha^T \mu_1 + \lambda_1 \alpha^T \Sigma_1 \alpha = \alpha^T \mu_2 - \lambda_2 \alpha^T \Sigma_2 \alpha \tag{13b}$$

Here, λ_1 and λ_2 are two scalars and resulting procedure to estimate these parameters is referred to as **minmax procedure**. The minmax procedure is an admissible procedure[18] when $\eta_1 = \eta_2$. For admissible procedure, $\lambda = \lambda_1 = 1 - \lambda_2$. If we substitute analytical characterization of α, β in η_1, η_2, the following equality must hold for admissible procedure.

$$\eta_1^2 - \eta_2^2 = \alpha^T [\lambda^2 \Sigma_1 - (1 - \lambda)^2 \Sigma_2] \alpha = 0 \tag{14}$$

The above criterion is a necessary condition to get the best approximation of amount of overlap. We will call λ *"overlap parameter"*, as it is a deciding factor which completely characterizes the overlap for a given first and second order Gaussian moments. Here, value of overlap parameter λ is determined heuristically using equation 14 as a base condition. Algorithm 1 is used to determine optimal value of λ. The overlap parameter λ completely dictates the linear separator parameters α and β. Once we get optimal value of λ, we can estimate linear separator parameters α, β. Since the procedure is admissible, we can compute $\mathbb{P}_1(\eta_1)$ = $\mathbb{P}_2(\eta_2)$ = $\mathbb{P}_{minmax}$. The amount of overlap(Υ) can be computed as below,

$$\eta_1 = h_1(\alpha, \beta), \quad \alpha = h_2(\lambda), \quad \beta = h_3(\lambda)$$
$$\Upsilon = \mathbb{P}_1(\eta_1) + \mathbb{P}_2(\eta_2) = \mathbf{2}\mathbb{P}_{minmax} = h(\mu_1, \mu_2, \Sigma_1, \Sigma_2, \lambda) \tag{15}$$

Here, we can notice that component of overlap(Υ) is parameterized by linear separator parameters α and β. Overlap parameter λ is completely dictating α and β as evident from equation 13. Hence, λ dictates the amount of overlap Υ. Here, λ is determined through an iterative procedure. As seen in figure 3, the overlap is determined by substituting for different values of λ in 15, a typical iterative routine settles for some value of λ, when the overlap is correctly determined, this can be noticed in Fig 3(d)-3(f). When we estimate area of overlap for the two sets of Gaussian distributions in figure 2 using algorithm 1, the area of overlap turns out to be Υ = 0.0706, which is same for both sets of Gaussian distributions shown in figures 2(a) and 2(b). Purpose of figure 3 is to empirically establish our finding that any two sets of Gaussian distributions touching at same confidence contours have same amount of overlap calculated according to algorithm 1. We exploit this finding to uniquely determine prescribed area of overlap based on user specified confidence contour threshold. In other words, there is a unique mapping between area of overlap of two k variate Gaussians and confidence contours c_t(where they are touching each other). The current scope of this paper explores this concept for $k = 2, 3$. We create a table that would give us unique value of overlap for a given value of c_t. We show relation between contour of touch(c_t) and area of overlap(Υ), in the figure 3(g)-3(h).

Algorithm 1 OverlapEstimation(μ_1, Σ_1, μ_2, Σ_2, *precision*)

```
initialize increment, criterion, λ
repeat
    calculate α with 13a
    calculate criterion with 14
    if criterion > precision
        then λ ← λ - increment
    if criterion < -precision
        then λ ← λ + increment
    increment ← increment/2
until -precision ≤ criterion ≤ precision
return λ
```

V. Probabilistic Collision Avoidance as Overlap between Two Gaussians

Equation 7 represents analytical expression of chance constraint defined in section IV. Our objective is to maximize the probability of collision avoidance. Here, we propose a novel formulation by modeling a non-linear chance constraint as overlap between component of two Gaussian distributions. We demonstrate in figure 4 that minimization of overlap

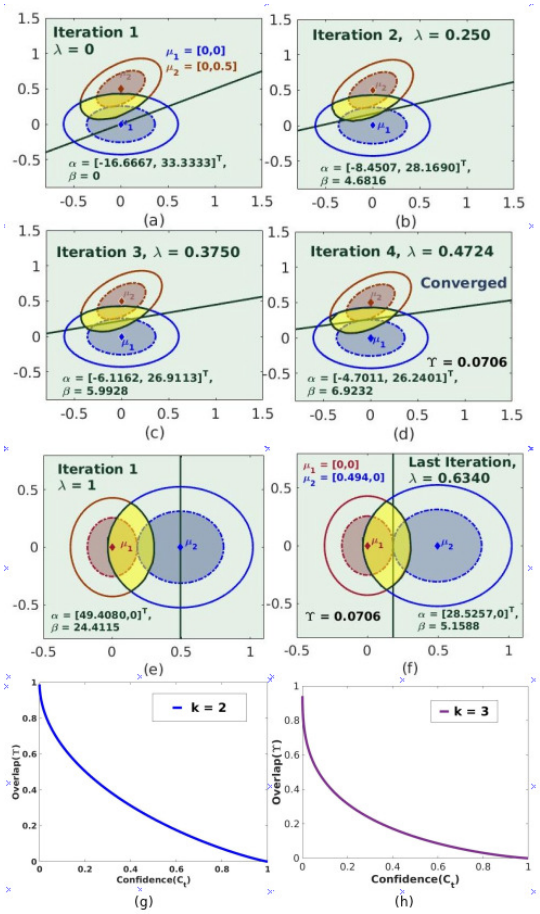


Fig. 3. This figure demonstrates how the values of overlap parameter λ evolves as algorithm 1 executes. It starts with initialized value of λ and ultimately converges to the value where the linear separator divides the two Gaussian distributions at the same confidence interval. From figure 3(a) -3(d), the Gaussian configuration considered is same as that of figure 2(a). Both the Gaussian distributions are touching at confidence contour corresponding to 80.51%. Equality constraint of equation 14 ensures that both the Gaussians are touching at the same confidence interval. The line drawn in the figure is $\alpha^T x = \beta$. We can see orientation and position of the line evolving as overlap parameter λ converges. For converged value of λ, we can see that the optimal linear separator passes through the point of touch of the two Gaussian distributions. Figure 3(e) -3(f) shows λ converging for Gaussian configuration considered in figure 2(b). We can see value of overlap turning out to be same for both sets as shown in figure 3(d)-3(f).

between two Gaussians is analogous to maximization of probability of collision avoidance.

A. Trajectory optimization with chance constraints

In section III, we discussed deterministic trajectory optimization routine. In this section, we reformulate this routine to accommodate state uncertainty of the agent and obstacle. We express collision avoidance constraint as desired measure of overlap between Gaussian populations of an agent and obstacle. The remodeling will use theory of overlapping of Gaussians described in section IV-C. Since there is an uncertainty in agent's position, we will redefine $J_{terminal}$ as below,

$$J_{terminal} = (\hat{X}_{t_N} - G_f)(\Sigma^d_{t_N})^{-1}(\hat{X}_{t_N} - G_f)^T \tag{16}$$

Equation 16 is Mahalanobis distance, which characterizes the number of standard deviations a point is away from mean of a distribution. Equation 16 will minimize number of standard deviations goal point is away from the mean position of the agent at the end of the trajectory.

Here, we are assuming no uncertainty in the actuation. Further, we assume that belief propagation for agent and obstacles for a given time-horizon is known. Hence, minimization of overlap between agent and obstacle populations can be thought of as minimum number of standard deviations an agent should deviate from its path to ensure collision free trajectory. While planning the trajectory, our constraint is to ensure that the minimum value of c_{t_i} for $i \in \{1, 2..., N\}$ is larger than certain threshold c_{min}, a user specified confidence contour which can't be penetrated. Which is analogous to saying that the overlap between an agent and obstacle at any time instant should not be greater than the prescribed overlap threshold Υ_{max}. c_{t_i} represents confidence contour percentage at which agent and obstacle/object are touching each other at time t_i.

1) Reformulation of collision avoidance constraint: Our goal is to reach G_f from the start position X_0 in N time-steps, each time-step of duration τ. Here, our objective is to find optimal set of velocity commands $V_{t_i} = (v^x_{t_i}, v^y_{t_i}, v^z_{t_i})$ which would satisfy our constraints as well as minimize the cost. We will be using the process model of the agent explained in section III.

Let the overlap between agent and obstacle j at time instance t_i be $\Upsilon^j_{t_i}$, which is dictated by overlap parameter $\lambda^j_{t_i}$. then,

$$\hat{X}_{t_i} = f(\hat{X}_{t_0}, V_{t_1}, V_{t_2}, V_{t_3}....V_{t_i}) \tag{17a}$$
$$\Upsilon^j_{t_i} = g_1(\hat{X}_{t_i}, \Sigma^d_{t_i}, \hat{O}^j_{t_i}, \Sigma^{o_j}_{t_i}, \lambda^j_{t_i}) \tag{17b}$$

Equation 17b is completely analogous to equation 15. In equation 17b, overlap is expressed in terms of agent/obstacle means, their corresponding uncertainties and overlap parameter. Mean position of the agent is expressed in terms of control commands$(V_{t_1}, V_{t_2}, ...V_{t_i})$ as explained in motion model(equation 4). Hence, equation 17b is completely parameterized by control commands$(V_{t_1}, V_{t_2}, ...V_{t_i})$ and overlap parameter$(\lambda^j_{t_i})$. Similarly, We can express condition to admissible procedure(equation 18a) in terms of control and overlap parameter as below,

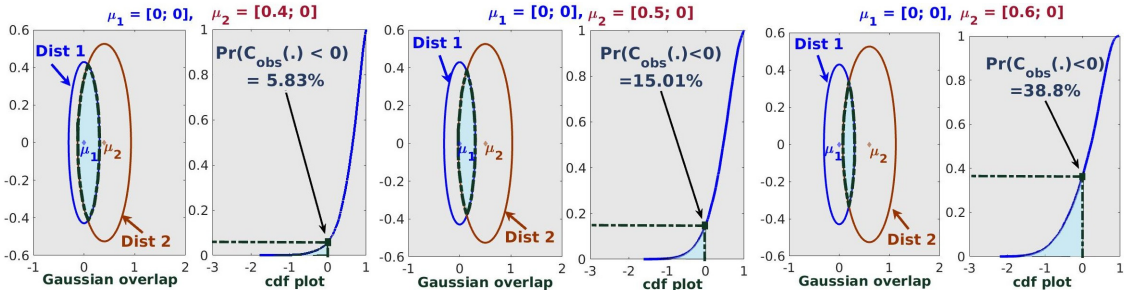
$$\left.\begin{aligned}(\eta^d_{t_i})^2 - (\eta^{o_j}_{t_i})^2 &= g_2(\hat{X}_{t_i}, \Sigma^d_{t_i}, \hat{O}^j_{t_i}, \Sigma^{o_j}_{t_i}, \lambda^j_{t_i}) \\ (\eta^d_{t_i})^2 - (\eta^{o_j}_{t_i})^2 &= f_1(\lambda^j_{t_i}, V_{t_1}, V_{t_2}, V_{t_3}...V_{t_i})\end{aligned}\right\} \tag{18a}$$
$$\Upsilon^j_{t_i} = 2\mathbb{P}^d_{t_i}(\eta^d_{t_i}) = 2\mathbb{P}^{o_j}_{t_i}(\eta^{o_j}_{t_i}) = f_2(\lambda^j_{t_i}, V_{t_1}, V_{t_2}...V_{t_i}) \tag{18b}$$

Equation 18a-18b are in accordance with equation 14-15, equation 18a models necessary condition for the procedure to be admissible in nature. It is important to note that in equations 17b and 18a, $\Sigma^d_{t_i}, \hat{O}^j_{t_i}$ and $\Sigma^{o_j}_{t_i}$ are known quantities. Hence, the only unknowns we are interested in solving are $\hat{X}_{t_i}$ and $\lambda^j_{t_i}$. Where, $\hat{X}_{t_i}$, mean of the agent at time t_i, which is a function of control commands according to motion model equation.

The maximum allowed overlap Υ_{max} is uniquely related to minimum number of standard deviations not allowed to be penetrated(c_{min}). c_{min} is expressed in terms of confidence intervals directly. So for a particular confidence interval, c_{min} will have a unique scalar value. Hence, a unique Υ_{max} value as explained in section IV-C. So, a chance constraint in terms of overlap between two Gaussians can be expressed as below.

$$C^{t_i}{}_{obs_j}(.) = \begin{cases}(\eta^d_{t_i})^2 - (\eta^{o_j}_{t_i})^2 = f^{lin}_1(\lambda^j_{t_1}, V_{t_1}, ..., V_{t_i}) = 0 \\ \Upsilon^j_{t_i} = f^{lin}_2(\lambda^j_{t_1}, V_{t_1}, ..., V_{t_i}) \leq \Upsilon_{max}\end{cases} \tag{19}$$
$$\forall\ i \in \{1, 2, 3, ...\ N\}\ ,\ \forall\ j \in \{1, 2, 3, ...\ P\}$$

Both the constraints of collision avoidance constraint(equation 19) are parameterized by velocity



(a) (b) (c)

Fig. 4. We establish an analogy between minimization of overlap between Gaussians as maximization of probability of collision avoidance modeled according to equation 6. Through an illustration, we demonstrate that as the amount of overlap between two distributions decreases, probability of collision avoidance increases. This is aptly conveyed through figure 4(a)-4(b)-4(c). For example, in figure 4(a), for a significant amount of overlap, the area below **0** in the cdf plot is very less. However, in figure 4(b) and 4(c), as the overlap decreases(achieved through incrementally separating μ_2 from μ_1), the corresponding area below **0** observed in the cdf plots significantly increases. Thus, it is clear that as the overlap between these two distributions decreases, the probability of collision avoidance[6](conveyed through cdfs) increases. The cdf plots of equation 6 were generated through *ecdf()* function of Matlab. The covariances considered for this demonstration are, $\Sigma_1 = \begin{pmatrix} 0.01 & 0 \\ 0 & 0.02 \end{pmatrix}$ and $\Sigma_2 = \begin{pmatrix} 0.03 & 0 \\ 0 & 0.03 \end{pmatrix}$.

controls($V_{t_1}, V_{t_2}, ...V_{t_i}$) and overlap parameter($\lambda^j_{t_i}$). $\Upsilon^j_{t_i} \leq \Upsilon_{max}$ ensures that the agent is avoiding the obstacle with certain minimum confidence. $(\eta^d_{t_i})^2 - (\eta^{o_j}_{t_i})^2 = 0$ ensures that the overlap parameter($\lambda^j_{t_i}$) we get through SCP routine is optimal and satisfies condition to admissible procedure. Closed form expressions for $f_1(.)$ and $f_2(.)$ are functions of optimization variables($V_{t_1}, V_{t_2}, ..., V_{t_N}, \lambda^j_{t_i}$) which are computed using mathematica[19]. $f_1(.)$ and $f_2(.)$ are non-linear in our variables of interest. We linearize them as shown below,

$$f_1^{lin}() = \bar{f}_1() + \sum_{k=1}^{i} \nabla_{V_{t_k}} (V_{t_k} - \bar{V}_{t_k}) + \nabla_{\lambda^j_{t_i}} (\lambda^j_{t_i} - \bar{\lambda}^j_{t_i}) \quad (20)$$

$$f_2^{lin}() = \bar{f}_2() + \sum_{k=1}^{i} \nabla_{V_{t_k}} (V_{t_k} - \bar{V}_{t_k}) + \nabla_{\lambda^j_{t_i}} (\lambda^j_{t_i} - \bar{\lambda}^j_{t_i}) \quad (21)$$

$$\forall\, i \in \{1, 2, 3, ... N\}\,, \forall\, j \in \{1, 2, 3, ... P\}$$

$f_1^{lin}(.)$ and $f_2^{lin}(.)$ are affine approximations of $f_1(.)$ and $f_2(.)$. $\nabla_{V_{t_k}}$ and $\nabla_{\lambda^j_{t_i}}$ are partial derivatives with respect to V_{t_k} and $\lambda^j_{t_i}$ respectively.

2) Trajectory optimization algorithm: We outline complete trajectory optimization algorithm built on above scheme. Let the trajectory of obstacle j be denoted by $\Omega_j = \{O^j_{t_1}, O^j_{t_2}, O^j_{t_3}, ...O^j_{t_N}\}$ and trajectory of all P obstacles be $\mathbf{\Pi_P} = \{\Omega_1, \Omega_2, \Omega_3, ...\Omega_P\}$. Let overlap parameter between obstacle j's distribution and agent's distribution for N timesteps be $\mathbf{\Lambda_j} = \{\lambda^j_{t_1}, \lambda^j_{t_2}...\lambda^j_{t_N}\}$, $j \in \{1, 2, 3, ... P\}$ and controls we are solving for be $\mathbf{V(t)} = \{V_{t_1}, V_{t_2}, V_{t_3}, ..., V_{t_N}\}$. Algorithm 2 outlines proposed SCP routine where we jointly solve for control($\mathbf{V(t)}$) and overlap parameter($\mathbf{\Lambda_j}$). In this algorithm, linearization control points $\mathbf{\bar{V}(t)} = \{\bar{V}_{t_1}, \bar{V}_{t_2}, \bar{V}_{t_3}, ..., \bar{V}_{t_N}\}$ are obtained by running the algorithm without collision avoidance constraints. That will give us an optimal trajectory from start to goal without any deviation. While overlap parameter linearization points are initialized with 0.5. Which implies that, $\mathbf{\bar{\Lambda}_j} = \{0.5, 0.5, ..., 0.5\}$. Subsequently, solution of previous iteration is used as linearization point in current iteration, which is implied in stage 4 and 5 of algorithm 2.

VI. Results and Discussions

We construct a model predictive control framework by using proposed probabilistic trajectory optimization routine as a base. Our proposal has been extensively evaluated for wide range of safety critical configurations. We show two such challenging situations to evaluate our proposal. We use Matlab based CVX[20] to prototype many of these scenarios. For a faster implementation, we use python based CVXOPT[21]. The simulations are carried out using Rotors[22], which is a micro aerial vehicle simulation framework built in gazebo[23]. The proposed approach is implemented in a model predictive control framework(MPC) along the lines of [24], [24] uses the idea of a receding horizon as the basis for building an MPC. We plan for a finite horizon and upto some intermediate way-point along the original trajectory.

Algorithm 2 ProbabilisticTrajOpt(Υ_{max}, Π_P, Σ_{agent}, $\Sigma_{obstacle}$)

1: **Initialization:** Guess for $\mathbf{\bar{\Lambda}^k_j(t)}$, $\mathbf{\bar{V}^k(t)}$, iteration counter $k = 0$
2: $\bar{D}^k(t) = Initialize Trajectory(\mathbf{\bar{V}^k(t)})$
3: **while** $\mid J_{k+1} - J_k \mid \geq \delta$ **do**

$$\mathbf{V^k(t)}, \mathbf{\Lambda^k_j(t)} = \text{argmin} \quad J_k$$

subject to

$$X_{t_i+1} = f(X_{t_i}, V_{t_i})$$
$$V_{min} \leq V_{t_i} \leq V_{max}$$
$$a_{min} \leq (\frac{V_{t_i+1} - V_{t_i}}{\tau}) \leq a_{max}$$
$$C_{obs_j}(\mathbf{\Lambda^k_j(t)}, \mathbf{V^k(t)}) \leq \mathbf{0},\ \forall\, j = \{1, 2, 3, ... P\}$$

4: $\mathbf{\bar{\Lambda}^k_j(t)} \leftarrow \mathbf{\Lambda^k_j(t)}$
5: $\mathbf{\bar{V}^k(t)} \leftarrow \mathbf{V^k(t)}$
6: $k \leftarrow k + 1$
7: **end while**

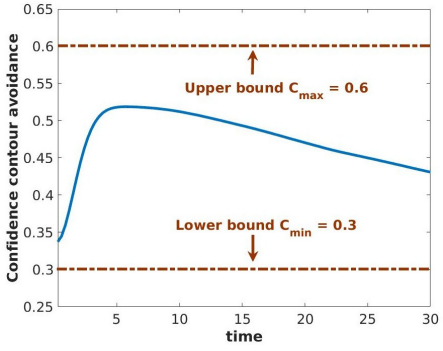


Fig. 5. This figure shows how confidence contour of touch($\mathbf{c_{t_i}}$) is changing over time for setting considered in VI-B. Throughout the trajectory, our agent is able to satisfy both, lower and upper-bound constraints. Initial rise in plot suggests that agent was slowing down first, and then it gradually accelerated to satisfy upper-bound constraint.

Intermediate way-points are placed at regular intervals to avoid extreme deviations in agent trajectory. In this section, we show results for two interesting applications. The detailed video of simulations under various safety critical situations is provided at ***https://robotics.iiit.ac.in/people/dhaivat.bhatt/video.html***

A. Multiple obstacle avoidance: Antipodal configuration

In this experiment, we show 3 obstacles attacking our agent in an antipodal configuration. The agent detects obstacles at sensing range of $S_r = 10$ meters. The acceleration bounds a_{min}, a_{max} are $-0.5\ m/s^2$, $0.5\ m/s^2$ respectively. Value of V_{min} and V_{max} are $0\ m/s$, $3.0\ m/s$ respectively. We keep a planning horizon of 28 steps, each of duration $\tau = 0.3$ seconds. We keep re-planning at every 0.3 seconds. Radius of agent/obstacles is taken as 0.5 meters. For this configuration, we consider $c_{min} = 90\%$ as our user specified confidence contour bound. In other words, our objective is to ensure that at least 90% confidence contours of the agent and obstacles do not penetrate each other during entire journey of the agent. As soon as it detects the obstacles, it starts deviating from its trajectory and 90% confidence contour of the agent avoids 90% confidence contours of all 3 obstacles. The position uncertainties considered are $\Sigma_{agent} = \Sigma_{obs1} = \Sigma_{obs2} = \Sigma_{obs3} = \begin{pmatrix} 0.02 & 0 & 0 \\ 0 & 0.02 & 0 \\ 0 & 0 & 0.02 \end{pmatrix}$. With this configuration, we show comprehensive results in figure 6. Sequence of images in figure 6 shows some snaps during obstacle avoidance maneuver, the navigation was successful and user specified lower bound was respected throughout the trajectory. As shown in confidence plots in figure 6(d)-6(e)-6(f), penetration for violet and dark green obstacles was at 95%, while for blue obstacle, it was 98%.

B. Object following in constrained corridor

We show another interesting application of our proposal. If agent and object of interest are entering in a constrained corridor, apart from putting lower bound c_{min}, we can also put upper bound(c_{max}) in such tight spaces. Putting upper bound will ensure that agent is not slowing down too much. Sub-constraint 2 of equation 19 will take a form like $\Upsilon_{min} \leq \Upsilon^j_{t_i} \leq \Upsilon_{max}$. Through demonstration, we advocate the usage of upper bound constraint in tight spaces or in situations where we need to follow/track object of interest. Upper bound ensures that agent is within certain range of object, which will reduce deviation in agent's trajectory, thus ensuring no collision with surrounding walls. We consider a case where agent and obstacle are entering in a constrained corridor at same time. In this case, user specified confidence contour bounds are $c_{min} = 30\%$ and $c_{max} = 60\%$. In other words, at least 30% confidence contours of agent and object can not penetrate into each other, while 60% confidence contours can't have 0 overlap at any time instant during the trajectory. Our planning horizon is of 40 timesteps, each of duration $\tau = 0.3$ seconds. Uncertainty being considered is, $\Sigma_{agent} = \begin{pmatrix} 0.02 & 0.01 & 0 \\ 0.01 & 0.02 & 0 \\ 0 & 0 & 0.02 \end{pmatrix}$ and $\Sigma_{object} = \begin{pmatrix} 0.03 & 0.02 & 0 \\ 0.02 & 0.03 & 0 \\ 0 & 0 & 0.02 \end{pmatrix}$. Before solving an MPC, these matrices are scaled up to incorporate radius of agent and obstacle. We consider non-isotropic uncertainty for this demonstration and show efficacy of our algorithm under tight spaces. In figure 7, We show snippets of various time-instances. The walls are modeled as stationary obstacles. For example, in figure 7(a), agent and object are entering in the corridor. Both upper bound and lower bound constraints are enforced and we can see that agent is able to maintain sufficient distance from the wall as well as the object while respecting the constraints. An absence of upper-bound constraint results in slowing down of the agent and we encounter a longer time for flight completion. Having upper bound favorably changes the velocity profile to complete trajectory in faster time. This shows usefulness of our proposal in tightly bounded spaces. We can use such modeling in object following. In figure 5, we show confidence plot for entire trajectory.

VII. Conclusion and Future Work

In this paper, a novel approach to dynamic collision avoidance and object following under uncertainty in state of robot and obstacles, modeled using Gaussian distribution, has been proposed. It has been derived by using area of overlap of the Gaussian distributions, which has unique characterization for a given confidence contour. The proposed algorithm, integrated with Linear MPC, optimizes over velocity profile and overlap parameter space to generate a navigation path in constrained dynamic environment. The paper puts forward results for two safety critical configurations. In future, we intend to model actuation dynamics into overlap of Gaussian framework and attempt to solve for challenging scenarios with unbounded co-variances. We can also model co-variance as a function of control and include into overlap of Gaussians framework. Another interesting task would be to extend this formulation for motion planning of non-holonomic vehicles. We can extend this formulation for multi-agent motion planning, where a centralized optimization routine will solve for control sequence of all agents with a guaranteed risk bound.

References

[1] M. Babu, R. R. Theerthala, A. K. Singh, B. Gopalakrishnan, and K. M. Kirshna, "Model predictive control for autonomous driving considering actuator dynamics," 2018.

[2] D. Kim, J. Kang, and K. Yi, "Control strategy for high-speed autonomous driving in structured road," in *Intelligent Transportation Systems (ITSC), 2011 14th International IEEE Conference on*. IEEE, 2011, pp. 186–191.

[3] C. Katrakazas, M. Quddus, W.-H. Chen, and L. Deka, "Real-time motion planning methods for autonomous on-road driving: State-of-the-art and future research directions," *Transportation Research Part C: Emerging Technologies*, vol. 60, pp. 416–442, 2015.

[4] W. Schwarting, J. Alonso-Mora, L. Pauli, S. Karaman, and D. Rus, "Parallel autonomy in automated vehicles: Safe motion generation with minimal intervention," in *Robotics and Automation (ICRA), 2017 IEEE International Conference on*. IEEE, 2017, pp. 1928–1935.

[5] G. Garimella, M. Sheckells, and M. Kobilarov, "Robust obstacle avoidance for aerial platforms using adaptive model predictive control," in *Robotics and Automation (ICRA), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5876–5882.

[6] J. Van Den Berg, P. Abbeel, and K. Goldberg, "Lqg-mp: Optimized path planning for robots with motion uncertainty and imperfect state information," *The International Journal of Robotics Research*, vol. 30, no. 7, pp. 895–913, 2011.

[7] I. Postlethwaite and M. Kothari, "Multi-agent motion planning for nonlinear gaussian systems," *International Journal of Control*, vol. 86, no. 11, pp. 2075–2089, 2013.

[8] R.-P. Lu, E. P. Smith, and I. Good, "Multivariate measures of similarity and niche overlap," *Theoretical Population Biology*, vol. 35, no. 1, pp. 1–21, 1989.

[9] B. Gopalakrishnan, A. K. Singh, M. Kaushik, K. M. Krishna, and D. Manocha, "Prvo: Probabilistic reciprocal velocity obstacle for multi robot navigation under uncertainty," 2017.

[10] B. Gopalakrishnan, A. K. Singh, and K. M. Krishna, "Closed form characterization of collision free velocities and confidence bounds for non-holonomic robots in uncertain dynamic environments," in *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*. IEEE, 2015, pp. 4961–4968.

[11] A. K. Singh and K. M. Krishna, "Reactive collision avoidance for multiple robots by non linear time scaling," in *Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on*. IEEE, 2013, pp. 952–958.

[12] L. Blackmore, H. Li, and B. Williams, "A probabilistic approach to optimal robust path planning with obstacles," in *American Control Conference, 2006*. IEEE, 2006, pp. 7–pp.

[13] M. Ono and B. C. Williams, "An efficient motion planning algorithm for stochastic dynamic systems with constraints on probability of failure." in *AAAI*, 2008, pp. 1376–1382.

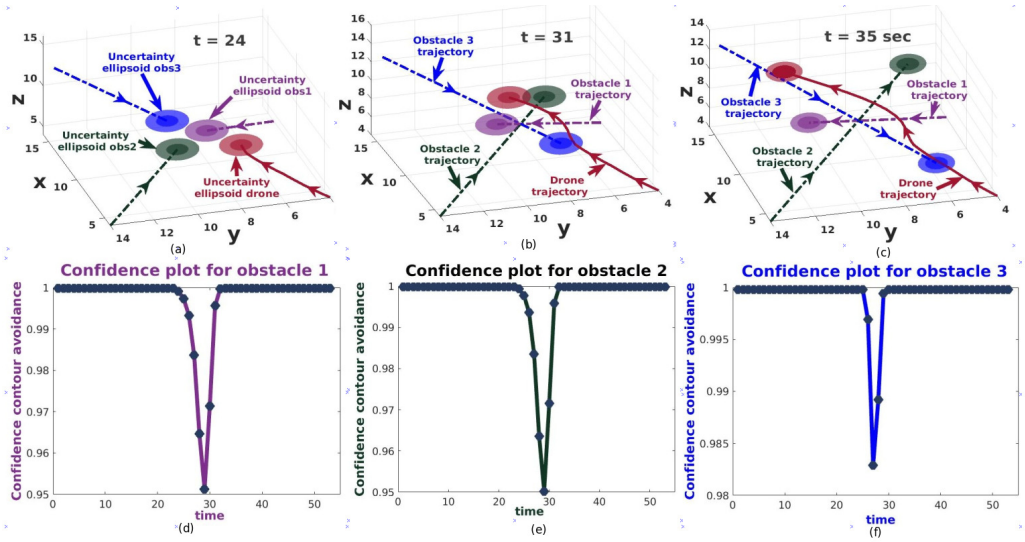


Fig. 6. Antipodal setting: The agent adopts a maneuver relating to a particular confidence of safety when it encounters obstacles in its sensing range. This is clearly shown in figure 6(a)-6(c). Figure 6(a), shows the situation, where the agent encounters obstacles, within its sensor range and starts taking appropriate control actions.Figure 6(b), highlights the resultant maneuver, that the agent adopts to achieve a targeted level of safety. Figure 6(c) shows the goal reaching ability of the agent, after avoiding obstacles. The lighter ellipsoidal shades in these figures represent the uncertainty region encompassing the mean positions of the agent and obstacles(filled with darker shades). Figures 6(d)-6(e)-6(f), shows the plots of confidence intervals for the collision avoidance maneuvers the that agent adopted in figures 6(a)-6(c). Our constraint was to ensure that 90% confidence contours of our agent avoids at least 90% confidence contours of all obstacles. From figures 6(d)-6(e)-6(f), we can observe that maximum overlap between the agent and any obstacle is 5%, i.e. only the 95% confidence contours grace each-other.

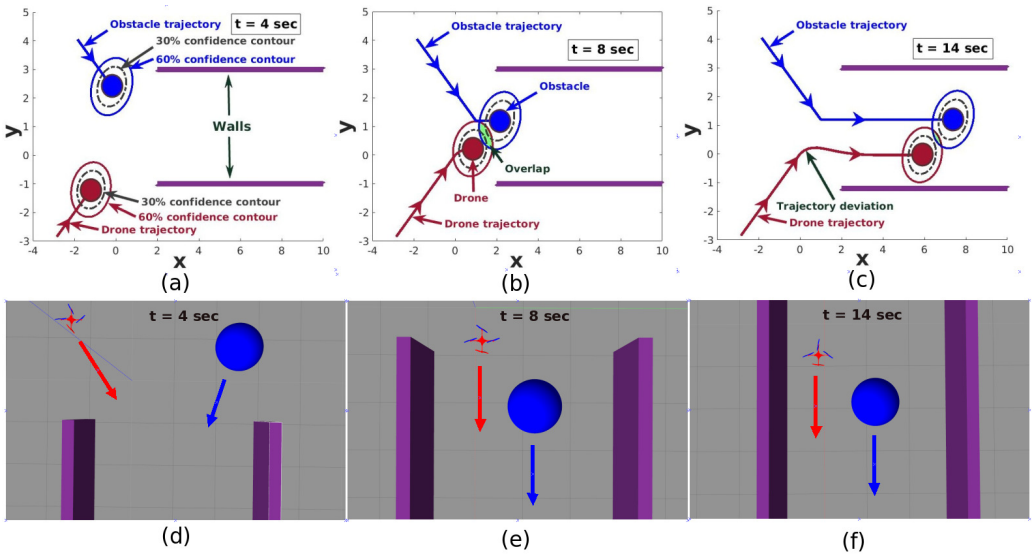


Fig. 7. Object following in constrained corridor: In figure 7(a), confidence contour for 30% and 60% are shown for both, agent and object. agent senses presence of obstacle at time $t = 4sec$. As agent enters in corridor, lower and upper bound constraints are enforced. Which ensures minimum risk behavior. For example, in figure 7(b)- 7(c), we can see overlap between agent and obstacle shaded in green for 60% confidence contours. While, 30% confidence contours which correspond to lower bound never penetrate each other as visible in figure 7(b)-7(c). Figure 7(e) shows top view of figure 7(b) in gazebo. Under such tightly bounded spaces, we observe that agent is able to safely maneuver constraints without crashing into walls. Trajectory deviation shown in 7(c) depicts that agent can maneuver in a way that would make it stay reasonably behind the object of interest while satisfying both lower and upper bound constraints. Figures 7(d)-7(e)-7(f) are gazebo results of figures 7(a)-7(b)-7(c) in bird's eye view.

[14] M. Ono, "Iterative risk allocation: A new approach to robust model predictive control with a joint chance constraint," in *Decision and Control, 2008. CDC 2008. 47th IEEE Conference on*, 2008, pp. 3427–3432.

[15] S.Boyd, "sequential convex programming," https://web.stanford.edu/class/ee364b/lectures/seq_slides.pdf, 2008.

[16] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bull. Calcutta Math. Soc.*, vol. 35, pp. 99–109, 1943.

[17] E. Nowakowska, J. Koronacki, and S. Lipovetsky, "Tractable measure of component overlap for gaussian mixture models," *arXiv preprint arXiv:1407.7172*, 2014.

[18] T. W. Anderson, R. R. Bahadur, *et al.*, "Classification into two multivariate normal distributions with different covariance matrices," *The annals of mathematical statistics*, vol. 33, no. 2, 1962.

[19] W. R. Inc., "Mathematica, Version 11.0," champaign, IL, 2016.

[20] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," http://cvxr.com/cvx, Mar. 2014.

[21] M. Andersen and L. Vandenberghe, "CVXOPT: A python package for convex optimization, version 1.1.9," http://cvxopt.org/index.html, 2016.

[22] F. Furrer, M. Burri, M. Achtelik, and R. Siegwart, "Rotorsa modular gazebo mav simulator framework," in *Robot Operating System (ROS)*. Springer, 2016, pp. 595–625.

[23] N. Koenig and A. Howard, "Design and use paradigms for gazebo, an open-source multi-robot simulator," in *Intelligent Robots and Systems, 2004.(IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on*, vol. 3. IEEE, 2004, pp. 2149–2154.

[24] A. Raemaekers, "Design of a model predictive controller to control uavs."

View publication stats

Applied Intelligence
https://doi.org/10.1007/s10489-019-01579-1

Quality based adaptive score fusion approach for multimodal biometric system

Keshav Gupta[1] · **Gurjit Singh Walia**[2] · **Kapil Sharma**[3]

© Springer Science+Business Media, LLC, part of Springer Nature 2019

Abstract

Multimodal Biometric Systems are extensively employed over unimodal counterparts for user authentication in the digital world. However, the application of multimodal systems to security-critical applications is limited mainly due to non-adaptiveness of these systems to the dynamic environment and inability to distinguish between spoofing attack and the noisy input image. In order to address these issues, a multimodal biometric system, which adaptively combines the scores from individual classifiers is proposed. For this, three modalities viz. face, finger, and iris are used to extract individual classifier scores. These classifier scores are adaptively fused considering that concurrent modalities are boosted and discordant modalities are suppressed. The conflicting belief among classifiers is resolved not only to achieve optimum fusion of classifier scores but also to cater dynamic environment. The proposed quality based score fusion also distinguish between spoofing attacks and noisy inputs as well. The performance of the proposed multimodal biometric system is experimentally validated using three chimeric multimodal databases. On an average, the proposed system achieves an accuracy of 99.5%, an EER of 0.5% and also outperforms state-of-the-art methods.

Keywords Image quality · Adaptive score fusion · Multimodal biometric

1 Introduction

Identity confirmation is very critical in today's digital world. Traditional pin and password methods are becoming obsolete day by day as it is easier to break them with an increase in technology. To address this, biometric systems which use people′s traits like face, fingerprints are considered a more appropriate solution in terms of robustness, reliability, and accuracy. Biometric systems using only a single modality like a fingerprint, face, etc. are known as unimodal systems. Unimodal systems possess different issues such as non-universality, intra-class variations, noise in input data, spoof attacks, and distinctiveness.[1]. These issues are limitedly handled by multimodal systems which combines information from multiple biometric modalities [1, 2]. However, with an increase in the number of identity frauds, a robust and reliable authentication system is required which can cater to security and privacy concerns.

In multimodal biometric systems, information obtained from multiple biometric traits is fused together using various techniques such as score level fusion, feature level fusion, and decision level fusion. [1, 2]. Fusion can be performed in serial and parallel modes. Researchers have worked on multimodal biometric systems at different fusion levels to show their efficiency over other traditional methods. For instance, Hossain et al. [3] proposed a serial fusion approach over face and iris modality. Here, multiple classifiers were arranged in a serial mode using 'best to worst' approach. On the other hand, Yang et al. fused the features of finger vein and fingerprint modalities to generate a more discriminative feature [4]. Although the feature set contains the highest amount of information, incompatibility

✉ Kapil Sharma
kapil@ieee.org

Keshav Gupta
keshavgupta101@gmail.com

Gurjit Singh Walia
gurjit.walia@gmail.com

[1] Department of Computer Science, Delhi Technological University, New Delhi 110042, India

[2] Defence Research and Development Organization, Ministry of Defence, New Delhi 110054, India

[3] Department of Information Technology, Delhi Technological University, New Delhi 110042, India

Published online: 17 December 2019

in type and dimension of features extracted from different modalities limits the application of feature level fusion. Generally, score level fusion is used to overcome the issues of feature level fusion. For example, Peng et al. proposed a t-norm based score level fusion method which combined various classifier scores to generate a final score [5]. The score fusion provided a trade-off between ease of use and efficiency. Fusion can also be performed at the decision level. For instance, Prabhakar and Jain worked in combining the results of four different fingerprint matching algorithms [6]. Fusion at the decision level is quite inflexible because of low information available and choice of the classifier. The most commonly used approach is fusion at score level due to low computation complexity and sufficient information content to discriminate [1, 2]. Biometric systems generally face challenges from spoof attacks where a fake biometric sample is used for illegitimate access. Apart from this, the computed match score highly depends on the quality of the input probe image.[7] Degraded environmental conditions or faulty devices used to capture the biometric feature may result in a poor match score for a genuine individual. This may result in poor performance of the biometric system.

To address the issue of spoofing attacks, Singh et al. [8] proposed an antispoofing technique for face recognition. The method presented identified face liveliness with morphological operations. For fingerprint spoof detection, Kho et al. [9] presented an incremental method based on SVMs and dynamic weight update procedure. Similarly, in order to address iris spoof attacks, Kaur et al. [10] presented an antispoofing method based on orthogonal features. Mostly, anti-spoof methods used feature specific properties of biometric traits to handle spoof attacks and may not be suitable for multimodal systems as it requires a more generic approach to handle spoof attacks.

While capturing input image, dynamic environmental conditions and other faulty feature-capture mechanisms introduce noise. This results in degradation of overall image quality. Poor quality images will generate a low match score and the biometric system's performance is degraded. Harin et al. [7] proposed a face recognition system to handle varying lighting conditions. Nathan et al. [11] proposed a unimodal biometric system using iris modality to deal with environmental issues as well as faulty device issues. Similarly, Pisani et al. [12] proposed an adaptive biometric system that deals with changing biometric feature over time, but a comprehensive solution which addresses all these issues simultaneously was not investigated. In order to resolve these issues, an adaptive multimodal system is required which not only handle poor environmental conditions like improper lighting or sweating fingers but also robust to spoofing attacks.

This paper presents an adaptive multimodal biometric system incorporating reliability factors for each modality. The reliability factor is estimated using image quality which can be a decisive factor in identifying fake biometric. The proposed technique addresses various issues simultaneously and works efficiently for multimodal systems. Few key points of the proposed method are summarised below.

- An adaptive multimodal biometric system using score fusion technique is proposed having three complimentary modalities namely fingerprint, iris and face. Multimodal systems highly decrease the chances of a successful spoof attack.
- Boosting of concurrent classifier scores and suppression of discordant classifier scores is performed simultaneously. This approach creates a clear and distinguished decision-boundary between an imposter and a genuine class.
- The reliability factor is calculated using no-reference quality measurement techniques for each modality. This not only adds to its adaptive nature but also makes it more robust under a dynamic environment and against spoof attacks.
- The proposed method is experimentally validated over three multimodal chimeric datasets generated using benchmarked images. The results depict high performance, low error rate and better detection of fake biometrics as compared to state of the art techniques.

The paper organization is as follows:

Related work in the field of the multimodal biometric system with a focus on the state-of-the-art score fusion methods are discussed in Section 2. The overview of the architecture of the proposed biometric system and details about the proposed score level fusion technique is discussed in Section 3. Section 4 contains the details of datasets used, qualitative and quantitative performance assessment of the proposed technique and exhaustive result analysis for the proposed biometric system. Section 5 draws the final conclusion and future directions.

2 Related work

Recently, multimodal biometric systems are extensively investigated for achieving robust and reliable solutions. In this section, we explicitly reviewed recent literature which is closely related to our work. Generally, biometric features can be combined at various levels such as feature level, score level and at the decision level. Combining similarity scores evaluated from different biometric traits is considered a suitable approach as it not only increases the reliability of the results but also reduces the overall complexity. There are various score fusion techniques proposed by researchers. For example, in [5], T-norms were used to fuse matching scores evaluated from multiple hand modalities like finger knuckle print, palm

print, finger vein, and fingerprint. Evaluated scores were also normalized and score fusion was performed using different T-norms. Similarly, Nanni et al. [13] used statistical and machine learning approaches for a combination of various fingerprint matchers on 4 FVC2006 databases. Authors investigated various score level fusion algorithms to check the best approach for score fusion and correlation among multiple fingerprint classifiers. In [14], authors proposed a multiple-instance score fusion using a finger-knuckle print of 4 different fingers. Match scores were first normalized and then fused together to reach a final decision. Also, Tao and Veldhuis [15] presented a score level fusion method using likelihood ratio under the assumption of Naïve Bayes. The likelihood ratio was estimated via operation points on ROC. Results were evaluated over the Face Recognition Grand Challenge (FRGC) 2D-3D face database. However, the traits used to distinguish between various users were limited to a single body part, these methods mostly suffered from universality problem.

To resolve the issue of universality, traits belonging to multiple regions were adopted. For instance, in [16], iris and facial features were fused at score level. In this, weights were assigned to individual scores from classifiers and a weighted score level fusion was performed. The fused score is used to take the final decision. Similarly, Sim et al. [17] presented a score fusion technique to combine similarity scores from the face and iris biometric traits. The experiments were performed on self-made "Universiti Teknologi Malaysia Iris and Face Multimodal Datasets" (UTMIFM) dataset along with the ORL face database and UBIRIS version 2.0 database. Also, Mukherjee et al. [18] proposed a novel fusion technique where different similarity scores were mapped to a single amalgamated match score for decision making. Parameters were tuned using differential evolution (DE) to reduce the overlapping of genuine and imposter score distribution area in a frequency distribution plot. Results were evaluated over two databases each having four different modalities viz. iris, fingerprint, left and right ear. Further, Liang et al. [19] presented a probabilistic score fusion algorithm which cast the fusion into an optimization problem having a natural order-preserving constraint. The effectiveness of the algorithm was demonstrated on two databases viz NIST-BSSR1 and XM2VTS-benchmark respectively. In [20], a hybrid approach using both score and decision level fusion was followed. Scores from individual classifiers were fused using Mean-closure weighting (MCW) and a decision was made based on DS theory over 3 virtual multimodal databases. In [21], a score fusion technique which combined scores from iris and face modalities was proposed. The authors deployed a fuzzy C-means clustering with level set (FCMLS) method to effectively localize iris images improving the overall results. Also, Liau and Isa [22] used support vector machines(SVM) to perform weighted score level fusion of optimized face and iris feature scores. However, the issue of conflicting match scores and the optimal combination was limitedly addressed among the techniques.

Nandakumar et al. [23] proposed an optimal match score fusion technique on the basis of a likelihood ratio test. A finite Gaussian mixture model was created using genuine and impostor score distribution. Experiments were conducted on 3 publicly available datasets NIST-BSSR1, XM2VTS-benchmark and WVU database. To resolve the conflict among classifier scores, Walia et al. [24] proposed a score fusion technique using PCR-6 with Backtracking Search Optimization. Similarly, Mezai et al. [25] presented a score fusion technique incorporating belief functions for iris and face modality. Authors used Denoeux and Appriou models to convert matching scores into belief assignments and PSO was used to compute the confidence factor. DS theory was then used to combine the masses and PCR-5 to predict the user's class. In [26] authors used graph diffusion technique to secure the biometric template and optimally fuse the individual classifier scores. Also, Kumar and Kumar [27] explored the ability of multimodal systems to adapt to the required security level. An Ant Colony Optimization (ACO) based parameter manipulates various other parameters like threshold, fusion technique, weights, etc. depending upon the given security level. However, the performance of these systems degraded in the presence of noisy input data.

Poh et al. [28] incorporated quality measures in multimodal biometric fusion which determined the reliability of the results given by fusion methods. In this, the quality information was used by Bayesian framework working with discriminative and generative classifiers to improve system's performance. Similarly, in [29], authors proposed a quality-dependent technique for score normalization in order to minimize the performance degradation which arises due to cross-device matching. Further, Shekhar et al. [30] investigated a joint sparse representation including a quality measure for each modality which optimized individual classifier scores. However, the performance for these methods was compromised under a dynamic environment and vulnerable against spoofing attacks.

Generally, multimodal biometric systems provide desired accuracy using fixed rules for combination and security level. But under dynamic conditions and ever-changing environment, the same rules may not be applicable or equally efficient. Keeping this issue in mind, an adaptive multimodal biometric system with score level fusion technique is proposed, which maps the matching scores into different domain by boosting or suppressing their values based on the threshold and security requirements to reach a final decision. The proposed method can

effectively distinguish between low-quality images and spoofing attacks. The next section presents a detailed overview of the proposed method.

3 Proposed multimodal biometric system

The architecture of the proposed multimodal biometric system is presented in Fig. 1. In this, three biometric features viz. Iris (i), Face (f), and Fingerprint (p) are fused using the proposed adaptive score level fusion.

Three biometric features are taken as input and corresponding features are extracted. For iris feature extraction, segmentation is done using an improvised hough transform method followed by normalization into rectangular blocks with fixed dimensions using Daugman's model. Finally, phase data extracted from 1-D log Gabor filter is quantized to encode unique pattern into a bit-wise biometric template. For extracting facial feature, Gabor filters are used which explore various visual properties like orientation selectivity, spatial localization, and spatial frequency characteristics. The feature vector is created by convolving the image with Gabor filters. For fingerprint trait, the input image is first enhanced using binarisation and thinning operations. Further, minutiae-based features are extracted from the corresponding image.

Comparison of the query image is performed with the templates stored in database and similarity match scores are obtained as S^i, S^f and S^p for iris, face and fingerprint respectively. These scores are processed and optimally combined using the proposed fusion model. Fusion model comprises of 3 stages: In the first stage, adaptive scores are calculated from the match scores for each modality. In the next stage, confidence and optimization factors are computed and finally score fusion is performed followed by normalization step. Finally, the fused score is compared with a threshold value to reach a decision. The proposed system is adaptive in nature as an adaptive score is calculated depending upon the distance of match score from a threshold value. Also, each modality is assigned with a reliability factor (α) using input image quality which provides unequal priors depending upon the reliability of input features. The following sub-section presents the details of the proposed Multimodal Biometric System.

3.1 Feature extraction and classifier score estimation

Multimodal modalities viz. iris, face, and fingerprint are processed for generic feature extraction and individual classifier scores are determined. For facial feature extraction, we adopted the Gabor filter approach for edge detection [31] due to low complexity, robustness against noise and other photometric disturbances [32]. This method recognizes a particular region of interest by capturing relevant frequency spectrum at specified orientations to extract features [31]. A Gaussian kernel function $\Upsilon_{\nu,\Theta}(x, y)$ is used to modulate

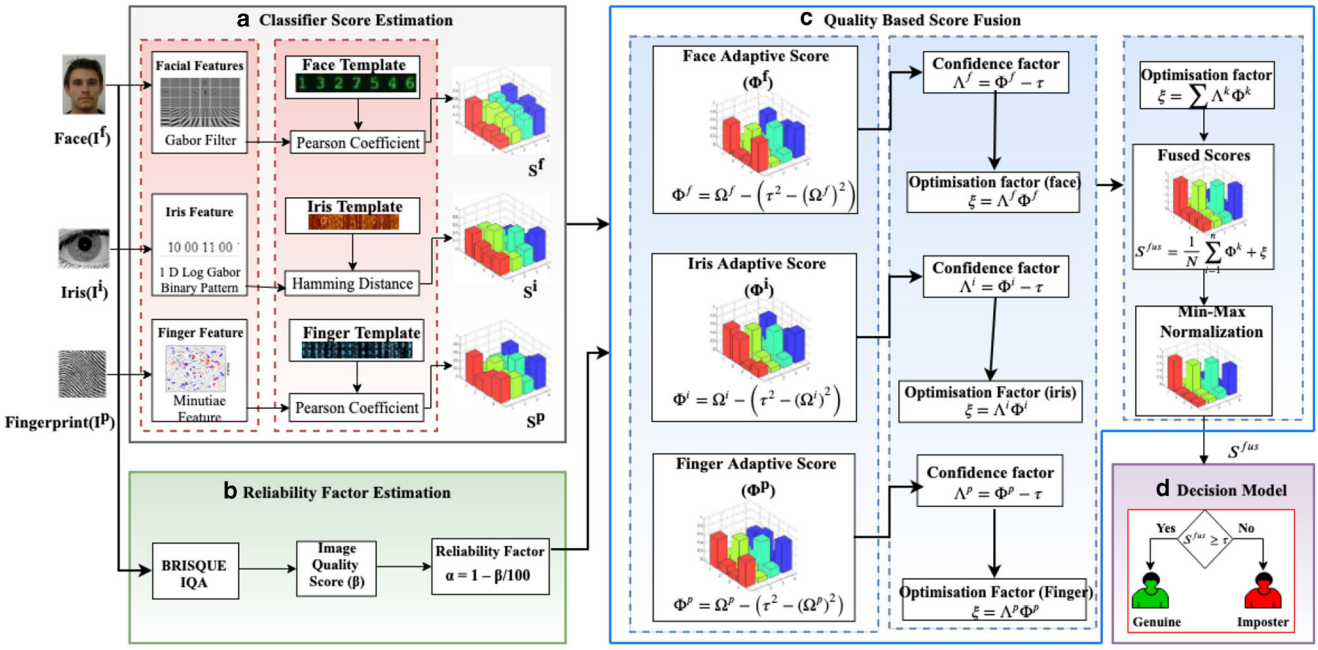


Fig. 1 Overview of the Proposed Multimodal Biometric System. Features from query image are extracted and compared with stored templates to generate individual classifier scores. Combined with reliability factor, scores are fused based on the proposed score fusion method to reach a final decision

the 2-D Gabor filter in form of a complex sinusoidal wave as in (1):

$$\Upsilon_{\nu,\Theta}(x, y) = \exp\left[-\frac{1}{2}\left\{\frac{x_{\Theta_n}^2}{\sigma_x^2} + \frac{y_{\Theta_n}^2}{\sigma_y^2}\right\}\right] \exp\left(2\pi \nu x_{\Theta_n}\right) \tag{1}$$

Where,

$$\begin{bmatrix} x_{\Theta_n} \\ y_{\Theta_n} \end{bmatrix} = \begin{bmatrix} sin\Theta_n & cos\Theta_n \\ -cos\Theta_n & sin\Theta_n \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \tag{2}$$

Here, ν is sinusoidal frequency, σ_x , σ_y are standard deviation along x and y direction of Gaussian envelop and Θ_n is the orientation defined in (3):

$$\Theta_n = \frac{\pi}{m}(n-1) \tag{3}$$

For n=1, 2... m where m represents the orientation count. Here, forty Gabor filters are used to convolve input grey facial image I^f in five scales and eight orientations followed by down-sampling by a factor of four to reduce redundancy before concatenating to form a feature vector, η^f which is stored in the database. Similarly, input facial probe image is convolved with Gabor filter bank to extract feature vector, ψ^f. The similarity match score between store template η^f and input probe image ψ^f is computed using Pearson's correlation coefficient using (4):

$$S^f = \frac{cov\left(\eta^f,\ \psi^f\right)}{\sigma_{\eta^f}\sigma_{\psi^f}} \tag{4}$$

where cov calculates the covariance between two vectors and σ represents their standard deviation.

For fingerprint feature extraction, a minutiae-based technique is employed. This technique is widely used by researchers [33, 34] for its high performance, low complexity and its analogy with methods used by forensic experts for fingerprint recognition. In this, input finger image I^p is first pre-processed through binarization and thinning. Binarization increases the contrast between ridges and valleys using (5):

$$B(m, n) = \left\{\begin{array}{ll} 1, & if\ I(m, n) \geq t \\ 0, & otherwise \end{array}\right\} \tag{5}$$

Where $I(m, n)$ represents the intensity value at pixel position (m, n) and t is threshold value. Thinning reduces ridges to unit-pixel thickness also known as skeletons and is performed using inbuilt morphological functions on binary images at Matlab platform. Minutiae are located over the thinned image using a 3x3 sliding window in a circular anti-clockwise manner to produce rutovitz crossing number (CN) [34] which defines the type of minutia and can be computed using (6):

$$CN = \frac{1}{2}\sum_{j=1}^{8}|q_j - q_{j-1}| \tag{6}$$

Where q_j represents pixel values of eight neighbors of any pixel q. Depending upon the value of CN, ridge pixel may be classified as isolated, ending, continuing, bifurcation and crossing point. Each minutia is then represented as a vector M=[m,n,CN,θ]. Where, (m, n) represents the coordinates of pixel p and θ is minutia orientation. For input finger image I^p, a feature template, η^p is generated by combining n minutiae using (7):

$$\eta^p = [M_1, M_2, \ldots M_n] \tag{7}$$

Similarly, a feature template, ψ^p for input probe image is generated. For computing the similarity match scores, minutiae are matched based on spatial distance and directional difference and a total number of matching minutiae are computed. Score S^p is computed between the acquired probe image and stored template using (8) as:

$$S^p = \frac{n_{match}^2}{n_\eta n_\psi} \tag{8}$$

Here, n_{match} represents the number of matching minutiae between two templates and n_η, n_ψ represents the total number of minutiae extracted.

For extracting the iris features, binary templates are generated using Khalil and Chadi [35] method as it improves the speed and accuracy of the iris segmentation process by accepting high quality images which also reduce the recognition error and produce a discriminating feature vector so as to improve the recognition accuracy and computational efficiency. In this, iris segmentation from input image I^i is performed using circular Hough transform [36] which provides center and radius of the iris. Further, the iris segment is normalized into a rectangular block with fixed dimensions using Daugman's rubber sheet model [37]. Additionally, localized iris texture is transformed from Cartesian to polar coordinates and iris texture is mapped in the radial direction using polar coordinates. The normalized iris is convolved with 1D Log-Gabor filter [38] whose frequency response is defined using (9):

$$G(\rho) = \exp\left\{-0.5 \times \frac{log\left(\frac{\rho}{\rho_0}\right)^2}{log\left(\frac{\sigma}{\rho_0}\right)^2}\right\} \tag{9}$$

Where ρ_0 represents central frequency and σ provides filter bandwidth. The extracted phase data from this convolution is quantized to four levels corresponding to four different phases. This results in a unique binary pattern, generating iris feature binary template η^i. Similarly, a feature template,

ψ^i for iris probe image is also generated. The similarity between the input query image and the stored template is calculated using hamming distance in (10).

$$HD\left(\eta^i,\ \psi^i\right) = \frac{1}{N}\sum_{j=1}^{n}\eta^i \bigotimes \psi^i \tag{10}$$

The hamming distance calculated between two templates is then converted into matching scores using the radial basis function (RBF) kernel in the range of [0, 1]. Using the RBF kernel, the match score between the input query image and the stored templates is computed as per (11):

$$S^i = exp\left(-\frac{HD\left(\eta^i, \psi^i\right)}{2\sigma^2}\right) \tag{11}$$

The calculated match scores from three modalities S^i, S^f, S^p are passed to the proposed fusion model to generate a fused score. The design of the proposed fusion model is discussed in the next section.

3.2 Quaity based adaptive score fusion

The fusion process is very important in a multimodal biometric system for making a decision. Here, we have proposed an adaptive score level fusion method with reliability factor (α) corresponding to each modality giving unequal priors depending upon the quality of input features. The proposed method performs boosting and suppression of individual classifier scores, which makes it adaptive under dynamic environment and robust against spoofing attacks.

Biometric image samples acquired under a dynamic environment may contain extra added noise. A reliability factor (α) based on image quality is calculated which provides a measure of reliability for each modality. Reliability factor (α) is estimated based on the No-reference quality assessment of input images. For this purpose, Blind/Referenceless image Spatial Quality Evaluator (BRISQUE) is adopted [39] which is used as an image quality metric. Here, Mean subtracted Contrast Normalized (MSCN) image is generated from the intensity image (I) using (12).

$$I'_k(x, y) = \frac{I_k(x, y) - \mu_k(x, y)}{\sigma_k(x, y) + 1} \tag{12}$$

where $k \in \{f, p, i\}$, (x,y) are spatial indices, μ and σ represents the mean and standard deviation respectively. Further, a Generalized Gaussian Distribution (GDD) is applied to obtain changes in coefficients distribution in the noisy image using (13).

$$f_k(x; p, \sigma^2) = \frac{p}{2q_k\Gamma(1/p)}\exp\left(-\left(\frac{|x|}{q}\right)^p\right) \tag{13}$$

where

$$q_k = \sigma\sqrt{\frac{\Gamma(1/p)}{\Gamma(3/p)}} \tag{14}$$

for $k \in \{f, p, i\}$, Γ is the gamma function, p is a shape parameter and σ^2 controls variance. Further, a brisque score is calculated using support vector regression (SVR) model trained on image database having similar distortions. The input image is compared to the SVR model with an RBF kernel providing a score value (β) in a range of 1-100. A low score value indicates a high quality of input image. Poor quality of the input query image suggests a high probability of the input image being fake or synthetic/reconstructed. In such cases, input biometric feature cannot be trusted and the reliability of biometric input image is reduced accordingly as per (15).

$$\alpha_k = 1 - \beta_k/100 \tag{15}$$

for $k \in \{f, p, i\}$, moreover, high quality input biometric feature results in high reliability. Thus, an overall reliability factor (α) is calculated for each biometric trait which denotes the reliability of each subject. This reliability factor (α) is incorporated with individual classifiers match scores (S^k) to generate optimized match scores with unequal priors using (16).

$$\Omega^k = \alpha_k * S^k \tag{16}$$

for $k \in \{f, p, i\}$ representing face, finger and iris modality. Thus, the reliability factor helps to tackle fake biometric features but also various dynamic environmental conditions where one modality is more reliable than any other modality by providing unequal priors. Individual classifier scores are optimized by calculating the adaptive scores (Φ^k) for each modality using (17) as:

$$\Phi^k = \Omega^k - \left(\tau^2 - \Omega^{k\,2}\right) \tag{17}$$

where, τ is an optimal threshold value and Ω^k denotes the match scores of individual classifiers for $k \in \{f, p, i\}$ representing the face, finger and iris modality. Further, a confidence factor (Λ) for each modality is calculated from the threshold value using (18) which indicates the score difference from the threshold value. Higher the difference, value of confidence factor will be high for both genuine and imposter scores.

$$\Lambda^k = \Phi^k - \tau \tag{18}$$

Using adaptive score, Φ_i and Confidence factor Λ_i, an Optimisation factor, ξ is computed for each modality as per (19)

$$\xi = \sum_{j=1}^{n}\Lambda^k\Phi^k \tag{19}$$

where n denotes the number of modalities. Optimization factor, ξ helps in determining the level of boosting or suppression to be done for individual match scores. The final fused score is estimated using (20) and is passed to a decision model for classification into genuine or imposter class.

$$S^{fus} = \frac{1}{N}\sum_{i=1}^{n} \Phi^k + \xi \tag{20}$$

The proposed score fusion is adaptive in nature as it performs boosting and suppression of individual classifier scores using (17) which helps in creating distinguished decision boundary for genuine and imposter class and robust against spoofing attacks as it incorporates quality based reliability factor using (16).

3.3 Decision model

The fused score is normalized using min-max approach and a final decision is performed using an optimal threshold value (τ) if the normalized fused score is greater than τ, then it is considered as Genuine else imposter. For validation of the proposed method, experiments are performed on three chimeric multimodal datasets generated using benchmark images. The next section provides the details of datasets used and its overall analysis.

4 Experimental validation

The performance of the proposed multimodal biometric system is evaluated over three multimodal databases in both qualitative and quantitative manner. In qualitative analysis, face, fingerprint and iris modalities are combined to generate a final score and are compared with individual classifier match scores. On the other hand, during quantitative analysis, various performance metrics like Decidability Index (DI), Equal Error Rate (EER), and Recognition Index (RI) are determined. In addition, the performance of the proposed score level fusion method is compared with other state-of-the-art methods.

4.1 Database & experimental design

We have obtained our Multimodal datasets from various benchmark datasets to validate our proposed algorithm. These Chimeric datasets are obtained by uniquely combining benchmark datasets namely CAS-PEAL Large-Scale Chinese Face Database [40], Casia-Face version 5.0 (Casia-FaceV5, http://biometrics.idealtest. org/), MCYT Bimodal Database [41], FVC2006 DB1-A fingerprint database [42], Casia iris database (Casia-IrisV1, http://biometrics.idealtest. org/), IITD PolyU iris database [43] and MMU2 iris database [44]. Few Sample images of the mentioned datasets are available in Fig. 2.

CAS-PEAL-R1 face database accommodates a total of 30,863 facial images from 1040 individuals out of which 595 are males and 445 females. The database is captured using 9 cameras to capture facial images with different poses, facial expressions, six accessories, and various lighting and background changes. Casia Face version 5.0 consists of 2500 16 bit color facial images having resolution 640*480 in BMP format of 500 distinct subjects. MCYT fingerprint database is obtained using two different sensors. These sensors namely CMOS-based capacitive capture device, and an optical capture device have a resolution of 500 dpi. Twelve samples with each sensor were captured

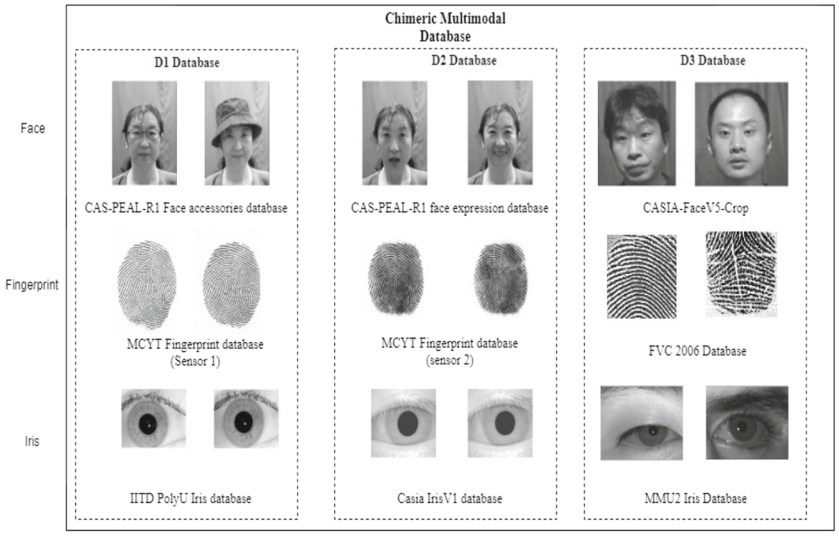


Fig. 2 Sample multimodal database images from the benchmark datasets

for each fingerprint from 330 subjects. Image resolution for captured images is 300x300 for sensor 1 and 256x400 for sensor 2 respectively. FVC2006 DB1-A contains 1680 uncompressed, 256 gray-levels fingerprint images from 140 subjects in BMP format. The images are acquired using an electric Field sensor with image size 96x96 having a resolution of 250 dpi. IITD PolyU iris database images are captured using a digital CMOS camera containing 5 samples from each eye of 224 subjects having a size of 320x 240 pixels. Casia IrisV1 database contains 756 iris images from 108 subjects in two different sessions. All captured images are in BMP format having a resolution of 320*280. MMU2 Iris database consists of 995 iris images in BMP format having resolution 320x238 pixels from 100 volunteers.

Three chimeric datasets are created namely D1, D2 and D3 to validate the proposed multimodal biometric system. D1 contains samples from N distinct subjects from CAS-PEAL-R1 accessories face database, MCYT fingerprint database from sensor 1 and IITD iris PolyU database. A virtual multimodal dataset containing N subjects is created by combining distinct subjects from each of the above-mentioned datasets. Similarly, D2 dataset is created from distinct subjects each from CAS-PEAL-R1 expression face database, MCYT fingerprint database from sensor 2 and Casia IrisV1 iris database and D3 dataset is created from distinct subjects each from Casia V5 face database, FVC2006 DB1-A fingerprint database, and MMU2 iris database. Also, all the subjects are different in D1, D2 and D3 database. In addition, a consolidated multimodal dataset, D4 of 3N subjects is also created by merging all subjects from D1, D2, and D3 databases. Five-fold cross-validation is performed with five different samples considering each sample as input subject once. We have implemented the proposed system on MATLAB 2018a platform with the hardware configuration of 4GB RAM and Intel i3 processor. The next section presents a qualitative analysis of the proposed system.

4.2 Performance metrics

The proposed system's performance is quantitatively analyzed by means of Decidability Index (DI), Equal Error Rate (EER), and Recognition Index(RI). Further, the results are compared with other state-of-the-art fusion methods. During this process, feature extraction and score calculation techniques are kept the same for evaluation of all the methods. Decidability Index (DI) is a performance metric used to quantify the distance between genuine and imposter score distributions and is calculated using (21):

$$DI = \frac{|\mu_g - \mu_i|}{\sqrt{(\sigma_g^2 - \sigma_i^2)/2}} \tag{21}$$

Here, μ_g , μ_i denotes the mean values, andσ_g and σ_i denotes the standard deviation values of genuine and imposter scores distributions, respectively. A high decidability index value suggests a higher ability of the classifier to separate genuine from imposters. Values of decidability index are shown in Table 1 for individual and proposed classifier over chimeric datasets D1, D2, and D3. Equal Error Rate (EER) is calculated by plotting ROC curves where the false acceptance rate (FAR) is plotted against the genuine acceptance rate (GAR). It provides the measure of the accuracy of the proposed biometric system. Recognition Index (RI) provides the recognition rate at rank-1 and can be used for performance evaluation. Cumulative Matching Characteristics (CMC) curves show the relationship between rank and the recognition rate.

4.3 Performance validation

The proposed adaptive multimodal biometric system is validated in a qualitative and quantitative manner. A qualitative analysis of the proposed system is presented in the next section.

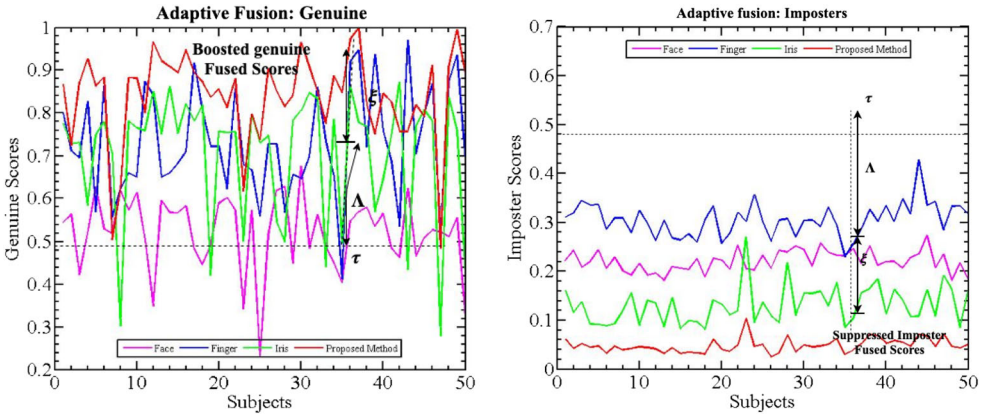
4.3.1 Qualitative validation

The performance of the proposed adaptive multimodal biometric system is validated over three multimodal databases and scores from individual classifiers are combined together using the proposed fusion method by boosting and suppression of the individual classifier match scores based on optimal threshold τ. For this, an adaptive score is calculated where scores above the optimal threshold are boosted to high values while scores below the threshold are suppressed to lower values. In the next step, a confidence factor (Λ^k) is calculated corresponding to adaptive match scores using (18) such that higher the value of score from the threshold, higher will be the corresponding confidence factor. Similarly, lower score values yield a low confidence factor corresponding to adaptive match scores. Further, an optimization factor is calculated using the confidence factor from individual classifiers. Boosting or suppression is carried out based on the value of the optimization factor which is decided on the basis of combined consensus from individual classifiers. The value of the optimization factor will be high if all three individual classifiers consider the subject to be genuine resulting in a boosted fused score. Similarly, if all individual classifiers consider a subject to be an imposter, a more suppressed fused score is generated as shown in Fig. 3.

Hence, score values change adaptively depending upon their distance from the threshold value. These two steps largely contribute to the boosting and suppression of individual classifier match scores and add to the adaptive

Table 1 Comparison of EER, DI, and RI values for D1, D2 and D3 databases

Method \ DB	D3 database			D4 database								
	EER	DI	RI	EER	DI	RI	EER	DI	RI	EER	DI	RI
Face	3.98±0.23	3.88±0.35	95±0.75	2.99±0.19	4.58±0.37	95±0.75	5.01±0.21	4.11 ±0.19	93±0.50	3.33±0.28	4.50 ±0.26	95±0.50
Finger	2.97±0.20	2.77±0.42	89±0.50	3.04±0.24	2.78±0.28	90±0.45	6.02±0.17	2.33±0.31	88±1.00	3.60±0.12	2.19±0.42	90±0.75
Iris	3.47±0.32	4.96±0.42	88±1.11	2.88±0.24	7.51±0.46	88±1.11	4.01±0.25	5.79±0.18	92 ±0.80	3.34 ±0.13	5.61±0.27	91±1.00
Frank t-norm	0.95±0.12	5.46±0.11	98±0.11	0.11±0.08	4.52±0.18	98±0.11	0.92±0.29	4.59±0.22	99±0.20	1.02±0.21	4.18±0.29	98±0.10
PCR5	0.79 ±0.26	5.13±0.16	98±0.31	0.20±0.15	7.98±0.21	99±0.15	0.22±0.17	7.67±0.14	98±0.40	0.69±0.16	6.83±0.15	98±0.20
PCR6 with BSA	0.99±0.22	5.78±0.28	98±0.51	0.78±0.22	7.82±0.18	98±0.35	0.87±0.16	7.56±0.17	99±0.10	0.91±0.23	5.80±0.25	98±0.25
PSO wtd. sum	1.99±0.27	5.17±0.29	97±0.51	0.89±0.27	6.88±0.19	98±0.55	1.06±0.19	6.23±0.12	98±0.45	1.47±0.29	5.61±0.29	98±0.50
Symmetric sum	2.00±0.31	4.93±0.35	95±0.47	1.00±0.23	7.20±0.18	98±0.55	1.14±0.19	4.17±0.13	96±0.75	1.64±0.17	5.12±0.25	96±0.60
Yager t-norm	4.0±0.36	2.62±0.21	90±0.41	4.50±0.31	2.23±0.33	90±0.50	3.50±0.38	2.64±0.25	91±0.60	4.21±0.27	2.57±0.43	90±1.00
Weighted Score Fusion	1.01±0.26	5.11±0.11	97 ±0.45	0.74±0.21	6.36±0.23	97 ±0.50	0.76±0.28	6.16±0.26	97 ±0.50	0.87±0.25	5.39±0.29	97±0.50
Fuzzy Score Fusion	1.08±0.21	4.92±0.24	99±0.50	0.88±0.31	4.18±0.23	95±0.40	2.99±0.27	3.52±0.14	96±0.50	1.65±0.18	2.95±0.34	97±0.45
Proposed Method	0.87±0.14	5.14±0.34	98±0.64	0.11±0.05	7.95±0.43	98±0.64	0.16±0.10	6.71±0.22	99±0.50	0.61±0.16	5.96±0.39	99±0.45



(a) Genuine Score values (b) Imposter Score values

Fig. 3 Comparision of Individual Classifier Scores with fused scores after boosting and suppression of (**a**) Genuine Scores and (**b**) Imposter scores

nature of the proposed biometric system. Boosting factor calculated from these two steps determines the amount of boosting and suppression to be performed and is used during the final score fusion process. This also leads to a higher separation between the peaks of genuine and imposter score distribution. The score distributions for individual classifiers as well as fusion model are presented in Fig. 4 Reliability factor (α) is also introduced corresponding to every individual input probe image depicting its reliability depending upon the environmental conditions, equipment used, etc. This reliability factor is computed using no-reference image quality scores provided by BRISQUE. If the input probe image possesses high noise, its quality will be low and vice-versa which helps in addressing various environmental challenges.

The frequency distribution of genuine and imposter scores of all 3 individual classifiers shows the smaller distance between peak values. This is generally due to irregularities in captured images, noise, and other environmental conditions. It basically represents a high rate of false acceptance and false rejection with smaller distance resulting in the decreased overall efficiency of the individual classifiers. On the other hand frequency distribution of scores for the proposed biometric system clearly shows the larger distance between Genuine and imposter classes and is also depicted in the quantitative analysis of results obtained.

4.3.2 Quantitative validation

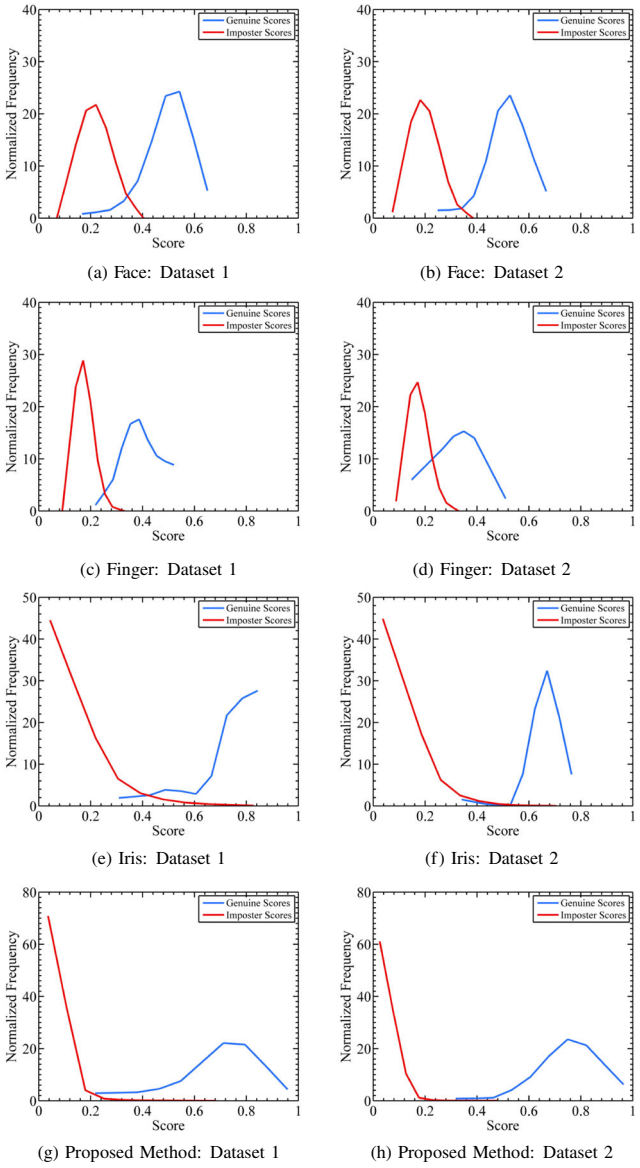
Quantitative analysis of the proposed system is performed on the basis of accuracy analysis, adaptivity analysis, and time-complexity analysis.

Accuracy Analysis The accuracy of the proposed system is analyzed using performance metrics namely EER, DI, and RI. The performance of the proposed method is compared with state-of-the-art techniques using these metrics as well. EER, DI, and RI values for various methods viz. T-norms (2011)[45], score fusion using PCR5 (2018) [46], score fusion using PCR6 with BSA (2019) [24], PSO weighted sum (2009) [47], Symmetric Sum (2018) [48], weighted score fusion (2014) [49] and fuzzy approach based score fusion (2016) [50] are compared with the proposed fusion method in Table 1

The EER value for proposed score level fusion method is 0.87 for D1 database, 0.11 for D2 database, 0.16 for D3 database and 0.61 for consolidated D4 database which is lowest in comparison with other state-of-the-art methods. The efficiency of the proposed system is also depicted by high decidability value of 5.14 for D1 database, 7.95 for D2 database, 6.71 for D3 database and 5.96 for D4 database and supported by ROC and CMC curves in Figs. 5 and 6 respectively. It is due to boosting and suppression of match scores which effectively distinguishes between Genuine and Imposter classes by creating a clear decision boundary. The variation among EER values is due to the use of different datasets and five-fold cross validation.

Quantitative analysis reveals that limitations of individual classifiers were effectively addressed by the proposed fusion method providing higher accuracy and more reliable results. The complimentary traits are fused together making system adaptive to dynamic environmental conditions and robust against spoofing attacks. The Roc curves in Fig. 5 and CMC curves in Fig. 6 shows the proposed fusion method performs better in comparison to single modality as well as other fusion methods.

Adaptivity Analysis To prove the adaptive nature of the proposed system, a new database was created by introducing extra noise to the original database which imitates acquired samples in a dynamic environment and spoofing attacks.



(a) Face: Dataset 1 (b) Face: Dataset 2

(c) Finger: Dataset 1 (d) Finger: Dataset 2

(e) Iris: Dataset 1 (f) Iris: Dataset 2

(g) Proposed Method: Dataset 1 (h) Proposed Method: Dataset 2

Fig. 4 Frequency Distribution Curves for D1 and D2 databases for (**a**) Face images over D1 (**b**) face images over D2 (**c**) Fingerprint images over D1(**d**) Fingerprint images over D2 (**e**) Iris images over D1 (**f**) Iris images over D2 (**g**) Proposed Method over D2 database (**h**) Proposed Method over D2 database

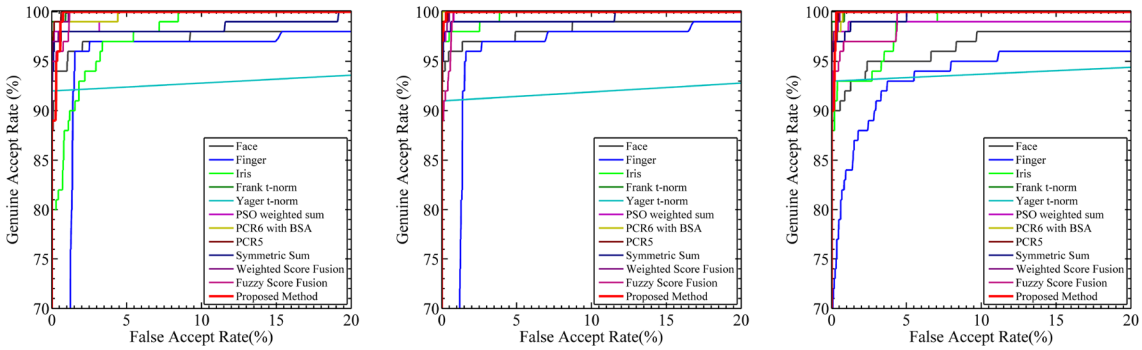
For this purpose, Gaussian noise with $\sigma = 0.05$ and offset $\delta = 0.01$ was introduced in iris and face images while for fingerprint images a Gaussian filter with a standard deviation of 0.4 is used. Sample images from the noise-induced database are shown in Fig. 7

Matching scores were calculated using noisy images as input probe images and the performance of the proposed biometric system was evaluated. During the fusion process, firstly, the noise was introduced in a single modality followed by noise introduction in two modalities and in the end noise was introduced in all three modalities and system's performance was measured in form of EER as tabulated in Table 2.

where $\bar{f}$, $\bar{p}$, $\bar{i}$ represents noisy modality for face, finger, and iris respectively. Low EER values indicate that the proposed system is able to give optimal performance when noise is introduced in one or two modalities. This represents dynamic environmental conditions where the acquired input probe image contains noise. The performance of system exponentially decreases when the input probe images from all three modalities were noisy. This represents a spoofing attack situation where synthetic/reconstructed samples with poor image quality are used. Thus the proposed multimodal biometric system is adaptive in nature withstanding dynamic environmental conditions and robust against spoofing attacks.

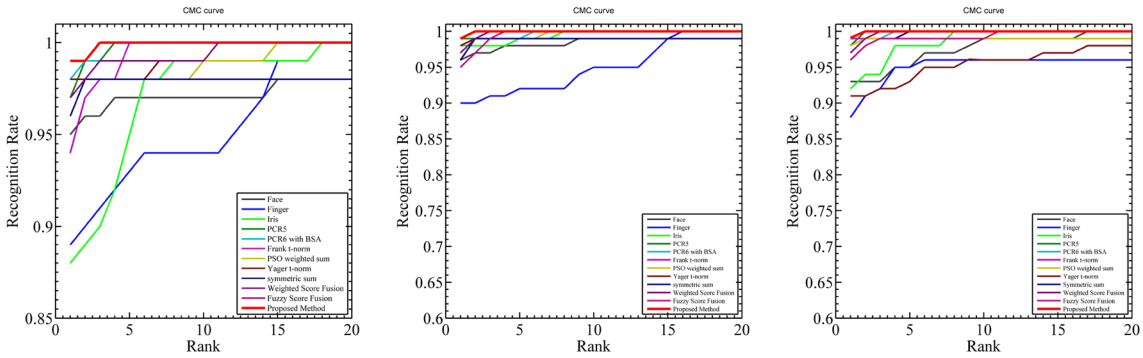
Time Complexity Analysis The computational efficiency of the proposed system is analyzed though time-complexity analysis of score fusion process of various methods. The time taken per subject by the proposed system is compared with other state-of-the-art methods in Table 3. It is evident from the results that the performance of the proposed method is comparable to other methods and better than its unimodal counterparts. The proposed method not only shows high performance but also provided added advantages after incorporating reliability factor.

In sum, the proposed fusion model optimally fuses complementary features making it adaptive to the dynamic environment and robust against spoofing attacks. Various facial distortion, iris off-angle blur, and insufficient boundary information in fingerprints limits the performance of individual classifiers and is also revealed in quantitative analysis. On the other hand, the proposed fusion method not only overcome these limitations by adaptively combining individual classifiers but also makes the system robust against spoofing attacks. An adaptive fusion is performed by boosting and suppression of match scores. Image quality is incorporated which boosts the adaptive nature of the proposed system under dynamic environment and robustness against spoofing attacks as compared with other state-of-the-art methods. The performance of the proposed method is validated with high decidability index, recognition index and low EER value are compared with other state-of-the-art methods. Moreover, lower time complexity makes the proposed system suitable for various real-time industry and security application. Hence, the proposed system shows improved performance along with adaptivity under dynamic environment and robust against spoofing attacks.



(a) D1 Database (b) D2 Database (c) D3 Database

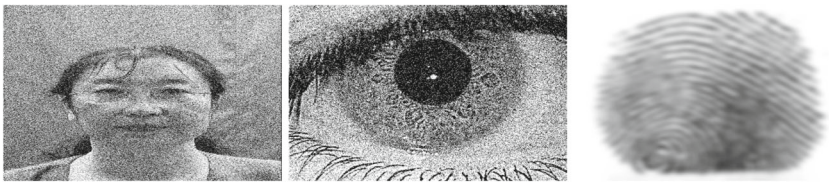
Fig. 5 Performance comparision of evaluated methods: ROC curves for D1, D2 and D3 database (**a**) ROC curves for various fusion techniques over Database D1 (**b**) ROC curves for various fusion techniques over Database D2 (**c**) ROC curves for various fusion techniques over Database D3



(a) D1 Database (b) D2 Database (c) D3 Database

Fig. 6 CMC curves for D1, D2 and D3 databases (**a**) comparison of CMC curves of various fusion techniques over Database D1 (**b**) comparison of CMC curves of various fusion techniques over Database D2 (**c**) comparison of CMC curves of various fusion techniques over Database D3

Fig. 7 Sample images from Noisy database



(a) Gaussian Noise (b) Gaussian Noise (c) Blurred

Table 2 Comparision of EER values for D1, D2 and D3 databases after adding Noise

Noise in modality	D1 database	D2 database	D3 database
$\bar{f}\,pi$	2.01±0.23	1.00±0.32	2.43±0.28
$f\,\bar{p}i$	1.80±0.17	0.15±0.21	0.91±0.15
$fp\bar{i}$	2.08±0.32	2.86±0.36	2.01±0.41
$\bar{f}\,\bar{p}i$	2.95±0.22	1.99±0.38	3.49±0.57
$\bar{f}\,p\bar{i}$	4.98±0.65	24.27±2.61	16.42±1.95
$f\,\bar{p}\bar{i}$	4.99±1.23	4.01±1.66	3.45±1.35
$\bar{f}\,\bar{p}\bar{i}$	39.85±4.72	58.07±5.81	60.78±5.08

Table 3 Comparision of time complexity for various Biometric recognition methods

Method	Time(ms)
Face	29.08
Finger	17.37
Iris	20.81
PCR5 [46]	17.43
PCR6 with BSA [24]	20.44
PSO wtd. Sum [47]	29.16
Symmetric Sum [48]	23.56
Frank t-norm [45]	20.26
Yager t-norm [45]	21.85
Weighted Score Fusion [49]	22.15
Fuzzy Score Fusion [50]	51.85
Proposed Method	20.56

5 Conclusion and future directions

This paper presents a novel adaptive score fusion technique for a multimodal biometric system using three biometric traits viz. fingerprint, face, and iris. The proposed technique performed boosting and suppression of individual scores from each modality. Reliability factor based on image quality is evaluated for each individual modality to resolve the problem of dynamic environment. It provides unequal prior to each classifier based on the quality of input images. The high value of the reliability factor improves the overall impact of the corresponding modality during score fusion. This not only helps in dealing with various problems of the dynamic environment but also very effective against spoofing attacks as well. The proposed system is evaluated on three chimeric multimodal databases generated from benchmark images of fingerprint, face and iris modality. Exhaustive result analysis shows that the proposed fusion technique overpass many state-of-the-art fusion methods under a dynamic environment. Quantitative analysis shows that the proposed adaptive score level fusion technique is computationally efficient for performing a fusion of scores from multiple classifiers. The qualitative analysis shows the distance between genuine and imposter score distribution is large resulting in high reliability and accuracy of the system.

In the future, incorporating user specific traits for estimating Reliability factor may also be used. This will result in an increase in the adaptive nature of the system and a more wide variety of challenges may be addressed. It may be done for each user or on the basis of the system's performance for a specific use. The proposed biometric system can also be customized to work at multiple security levels as per requirements.

References

1. Ross A, Jain A (2003) Information fusion in biometrics. Pattern Recogn Lett 24(13):2115–2125
2. Unar JA, Seng WC, Abbasi A (2014) A review of biometric technology along with trends and prospects. Pattern Recogn 47(8):2673–2688
3. Hossain M, Chen J, Rahman K (2018) On enhancing serial fusion based multi-biometric verification system. Appl Intell 48(12):4824–4833
4. Yang J, Zhang X (2012) Feature-level fusion of fingerprint and finger-vein for personal identification. Pattern Recogn Lett 33(5):623–628
5. Peng J, El-Latif AAA, Li Q, Niu X (2014) Multimodal biometric authentication based on score level fusion of finger biometrics. Optik 125(23):6891–6897
6. Prabhakar S, Jain A (2002) Decision-level fusion in fingerprint verification. Pattern Recogn 35(4):861–874
7. Sellahewa H, Jassim SA (2010) Image-Quality-Based Adaptive face recognition. IEEE Trans Instrum Meas 59(4):805–813
8. ManminderSingh AS (2017) Arora a robust anti-spoofing technique for face liveness detection with morphological operations. Optik 139:347–354
9. Kho JB, Lee W, Choi H, Kim J (2019) An incremental learning method for spoof fingerprint detection. Expert Syst Appl 116:52–64
10. Kaur B, Singh S, Kumar J (2019) Cross-sensor iris spoofing detection using orthogonal features. Comput Electr Eng 73:279–288
11. Kalka ND, Zuo J, Schmid NA, Cukic B (2010) Estimating and fusing quality factors for iris biometric images. IEEE Trans Syst Man Cybern - Part A: Syst Hum 40(3):509–524
12. Pisani PH, Lorena AC, de Carvalho AC (2018) Adaptive Biometric Systems using Ensembles. IEEE Intell Syst 33(2):19–28
13. Nanni L, Lumini A, Ferrara M, Cappelli R (2015) Combining biometric matchers by means of machine learning and statistical approaches. Neurocomputing 149:526–535
14. Shariatmadar ZS, Faez K (2014) Finger-knuckle-print recognition performance improvement via multi-instance fusion at the score level. Optik - Int J Light Electron Opt 125(3):908–910
15. Tao Q, Veldhuisl R (2013) Robust biometric score fusion by naive likelihood ratio via receiver operating characteristics. IEEE Trans Inf Forensic Secur 8(2):305–313
16. Sim HM, Asmuni H, Hassan R, Othman RM (2014) Multimodal biometrics: Weighted score level fusion based on non-ideal iris and face images. Expert Syst Appl 41(11):5390–5404
17. Sim HM, Asmuni H, Hassan R, Othman RM (2014) Multimodal biometrics: Weighted score level fusion based on non-ideal iris and face images. Expert Syst Appl 41(11):5390–5404
18. Mukherjee S, Pal K, Majumder BP, Saha C, Panigrahi BK, Das S (2014) Differential evolution based score level fusion for multimodal biometric systems. IEEE Symposium on Computational Intelligence in Biometrics and Identity Management (CIBIM), pp 1–7
19. Liang Y, Ding X, Liu C, Xue J-H (2016) Combining multiple biometric traits with an order-preserving score fusion algorithm. Neurocomputing 171:252–261
20. Dwivedi R, Dey S (2019) A novel hybrid score level and decision level fusion scheme for cancelable multi-biometric verification. Appl Intell 49(3):1016–1035
21. Roy K, Shelton J, O'Connor B, Kamel MS (2015) Multibiometric system using fuzzy level set, and genetic and evolutionary feature extraction. IET Biom 4(3):151–161

22. Liau HF, Isa D (2011) Feature selection for support vector machine-based face-iris multimodal biometric system. Expert Syst Appl 38(9):11105–11111
23. Nandakumar K, Chen Y, Dass SC, Jain A (2008) Likelihood Ratio-Based biometric score fusion. IEEE Trans Pattern Anal Mach Intell 30(2):342–347
24. Walia GS, Singh T, Singh K, Verma N (2019) Robust Multimodal Biometric System based on Optimal Score Level Fusion Model. Expert Syst Appl 116:364–376
25. Mezai L, Hachouf F (2015) Score-Level Fusion of face and voice using particle swarm optimization and belief functions. IEEE Trans Human-Mach Syst 45(6):761–772
26. Walia GS, Rishi S, Asthana R, Kumar A, Gupta A (2019) Secure multimodal biometric system based on diffused graphs and optimal score fusion. IET Biom 8(4):231–242
27. Kumar A, Kumar A (2016) Adaptive management of multimodal biometrics fusion using ant colony optimization. Inf Fusion 32(Part B):49–63
28. Poh N, Kittler J (2012) A unified framework for biometric expert fusion incorporating quality measures. IEEE Trans Pattern Anal Mach Intell 34(1):3–18
29. Poh N, Kittler J, Bourlai T (2010) Quality-Based Score normalization with device qualitative information for multimodal biometric fusion. IEEE Trans Syst Man Cybern - Part A: Syst Hum 40(3):539–554
30. Shekhar S, Patel VM, Nasrabadi NM, Chellappa R (2014) Joint sparse representation for robust multimodal biometrics recognition. IEEE Trans Pattern Anal Mach Intell 36(1):113–126
31. Liu C, Wechsler H (2002) Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. IEEE Trans Image Process 11(4):467–476
32. Abhishree TM, Latha J, Manikantan K, Ramachandran S (2015) Face Recognition Using Gabor Filter Based Feature Extraction with Anisotropic Diffusion as a Pre-processing Technique. Procedia Comput Sci 45:312–321
33. Farina A, Kovács-vajna ZM, Leone A (1999) Fingerprint minutiae extraction from skeletonized binary images. Pattern Recogn 32(5):877–889
34. Sudiro SA, Paindavoine M, Kusuma TM (2007) Simple fingerprint minutiae extraction algorithm using crossing number on valley structure. IEEE Workshop on Automatic Identification Advanced Technologies, pp 41–44
35. Kahlil AT, Chadi FEMA (2010) Generation of iris codes using 1D Log-Gabor filter. IEEE International Conference on Computer Engineering & Systems, pp 329–336
36. Ye H, Shang G, Wang L, Zheng M (2015) A new method based on hough transform for quick line and circle detection. IEEE International Conference on Biomedical Engineering and Informatics (BMEI), pp 52–56
37. Daugman J (2004) How iris recognition works. IEEE Trans Circ Syst Video Technol 14(1):21–30
38. Field D (1987) Relations between the statistics of natural images and the response properties of cortical cells. J Opt Soc Amer 4(12):2379–2394
39. Mittal A, Moorthy AK, Bovik AC (2012) No-Reference Image quality assessment in the spatial domain. IEEE Trans Image Process 21(12):4695–4708
40. Gao W, Cao B, Shan S, Chen X, Zhou D, Zhang X, Zhao D (2008) The CAS-PEAL Large-Scale chinese face database and baseline evaluations. IEEE Trans Syst Man Cybern - Part A: Syst Hum 38(1):149–161
41. Ortega-Garcia J, Fierrez-Aguilar J, Simon D, Gonzalez J, Faundez-Zanuy M, Espinosa V, Satue A, Hernaez I, Igarza J-J, Vivaracho C, Escudero D, Moro Q-I (2003) MCYT Baseline corpus: a bimodal biometric database. IEEE Proc - Vis Image Signal Process 150(6):395–401
42. Cappelli R, Ferrara M, Franco A, Maltoni D (2007) Fingerprint verification competition 2006. Biom Technol Today 15:7–9
43. Kumar A, Passi A (2010) Comparison and combination of iris matchers for reliable personal authentication. Pattern Recogn 43(3):1016–1026
44. MMU2 Iris Database[Online], Available: http://pesona.mmu.edu.my/ccteo/, Accessed: June, 2019
45. Hanmandlu M, Grover J, Gureja A, Gupta HM (2011) Score level fusion of multimodal biometrics using triangular norms. Pattern Recogn Lett 32(14):1843–1850
46. Sharma R, Das S, Joshi P (2018) Score-level fusion using generalized extreme value distribution and DSmt for multibiometric systems. IET Biom 7(5):474–481
47. Srinivas N, Veeramachaneni K, Osadciw LA (2009) Fusing correlated data from multiple classifiers for improved biometric verification. 12th International Conference on Information Fusion, Seattle, USA, pp 1504–1511
48. Cheniti M, Boukezzoula N-E, Akhtar Z (2018) Symmetric sum-based biometric score fusion. IET Biom 7(5):391–395
49. Sim HM, Asmuni H, Hassan R, Othman RM (2014) Multimodal biometrics: Weighted score level fusion based on non-ideal iris and face images. Expert Syst Appl 41(11):5390–5404
50. Fakhar K, Aroussi ME, Saidi MN, Aboutajdine D (2016) Fuzzy pattern recognition-based approach to biometric score fusion problem. Fuzzy Sets Syst 305:149–159

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Journal of Classical Analysis
Volume 14, Number 2 (2019), 137–153
doi:10.7153/jca-2019-14-11

RATE OF CONVERGENCE OF GUPTA–SRIVASTAVA OPERATORS BASED ON CERTAIN PARAMETERS

RAM PRATAP AND NAOKANT DEO*

Abstract. In the present paper, we consider the Bézier variant of the Gupta-Srivastava operators [14] and discuss some direct convergence results by using of Lipschitz type spaces, Ditzian-Totik modulus of smoothness, weighted modulus of continuity and for functions whose derivatives are of bounded variation. In the end some graphical representation for comparison with other variants have been presented.

1. Introduction

In the year 2003, Srivastava et al. [15] proposed a sequence of positive linear operators, containing some well-known operators as special cases and they studied the convergence properties. After a gap of two year Ispir and Yüksel proposed an important Bézier variant of Srivastava-Gupta operators and discussed the rate of convergence for the functions of bunded variation in the interval $[0,\infty)$. Several researchers have proposed different generalizations of the Srivastava-Gupta operators (see for instance [1], [4], [10], [11], [16], [17], etc.). One of them, Yadav [17] studied a modification of Srivastava-Gupta operators which preserve the constant functions as well as linear functions, established the Voronovskaya type theorem and statistical convergence. In 2017, Neer et al. [12] proposed the Bézier variant of the operators, which were introduced by Yadav [17] and discussed several convergence properties.

Recently, Gupta and Srivastava [7] proposed a general family of a positive linear operator, which preserve constant functions as well as linear functions for all $c \in \mathbb{N} \cup \{0\} \cup \{-1\}$. They have considered the general sequence of positive linear operators containing some well-known operators as special cases. For $m \in \mathbb{Z}$, and $c \in \mathbb{N} \cup \{0\} \cup \{-1\}$ operators defined in [7] are given as:

$$L^{c}_{n,m}(f(t);x) = \{n+(m+1)c\} \sum_{k=1}^{\infty} p_{n+mc,k}(x;c) \int_{0}^{\infty} p_{n+(m+2)c,k-1}(t;c) f(t)dt + p_{n+mc,0}(x;c) f(0), \tag{1}$$

Mathematics subject classification (2010): 41A25, 41A36.
Keywords and phrases: Baskakov operators, hypergeometric function, Lipschitz type space, Ditzian-Totik modulus of smoothness, functions of bounded variation.

* Corresponding author.

© ELEMENT, Zagreb

where

$$p_{n,k}(x;c)=\frac{(-x)^k}{k!}\phi_{n,c}^{(k)}(x)$$

and

$$\phi_{n,c}^{(k)}(x)=\begin{cases}e^{-nx}, & c=0,\\ (1-x)^n, & c=-1,\\ (1+cx)^{\frac{-n}{c}}, & c=1,2,3,\ldots.\end{cases}$$

Very recently, Pratap and Deo[14] studied some approximation properties of these operators and termed these operators as Gupta-Srivastava operators.

Inspired from the recent work depending upon some parameter $\alpha\geqslant 1$, we propose here the following sequence of operators (also called Bézier variant) (1) as follows:

$$\begin{aligned}F_{n,m}^{c,\alpha}(f(t);x)=&\{n+(m+1)c\}\sum_{k=1}^{\infty}Q_{n+mc}^{(\alpha)}(x;c)\int_0^{\infty}p_{n+(m+2)c,k-1}(t;c)f(t)dt\\&+Q_{n+mc,0}^{(\alpha)}(x;c)f(0),\end{aligned}\tag{2}$$

where $Q_{n+mc,k}^{(\alpha)}(x;c)=\left(J_{n+mc,k}(x,c)\right)^{\alpha}-\left(J_{n+mc,k+1}(x,c)\right)^{\alpha}$, $\alpha\geqslant 1$, with $J_{n+mc,k}(x,c)=\sum_{j=k}^{\infty}p_{n+mc,k}(x,c)$, where $k<\infty$ and otherwise zero. It is obvious that the operators $F_{n,m}^{c,\alpha}(.;x)$ are the linear positive operator. For $\alpha=1$ the operators (2) immediately reduce to the form (1).

The special cases of operators (2) are given below:

(i) For $c=0$, $\alpha=1$ and $\phi_{n,0}(x)=e^{-nx}$, we get Phillips operators

$$L_{n,m}^{0}(f(t);x)=n\sum_{k=1}^{\infty}p_{n+mc,k}(x;0)\int_0^{\infty}p_{n+mc,k-1}(t;0)f(t)dt+p_{n,0}(x;0)f(0),$$

where

$$p_{n,k}(x,0)=\frac{e^{-nx}(nx)^k}{k!}\ \text{ and }\ x\in[0,\infty).$$

(ii) For $c\in\mathbb{N}$, $\alpha=1$ and $\phi_{n,c}(x)=(1+cx)^{-\frac{n}{c}}$, we get genuine Baskakov-Durrmeyer type operators. These operators are similar to (1) except for $c=\{0,-1\}$, called summation integral type of operators, where

$$p_{n,k}(x;c)=\frac{\left(\frac{n}{c}\right)_k}{k!}\frac{(cx)_k}{(1+cx)^{\frac{n}{c}+k}}$$

and $(n)_i$ denotes the rising factorial given by

$$(n)_i=n(n+1)(n+2)...(n+i-1)\ \&\ (n)_0=1(i\in\mathbb{N}).$$

(iii) For $c=-1$, $\alpha=1$ and $\phi_{n,-1}(x)=(1-x)^n$, we have a sequence of Bernstein-Durrmeyer operators

$$L_{n,m}^{-1}(f,x)=(n-m-1)\sum_{k=1}^{n-m-1}p_{n-m,k}(x,-1)\int_0^1 p_{n-m-2,k-1}(t,-1)f(t)dt + p_{n-m,0}(x,-1)f(0)+p_{n-m,n-m}(x,-1), \tag{3}$$

where

$$p_{n,k}(x;-1)=\binom{n}{k}x^k(1-x)^{n-k}.$$

The purpose of this article is to investigate the approximation results by using of Lipchitz type space, Ditzian-Totik modulus of smoothness, weighted modulus of continuity and functions of bounded variation.

2. Basic results

In this section, we give some auxiliary results and with help of these results we study our main results.

Let $C[0,\infty)$ denotes the space of all continuous functions in $[0,\infty)$ and let $C_B[0,\infty)$ be the space of all continuous and bounded functions in $[0,\infty)$.

LEMMA 1. *Let* $f(t)=e_i=t^i$, $i=0,1,2,3,4$, $c\in\mathbb{N}\cup\{0\}\cup\{-1\}$ *and* $m\in\mathbb{Z}$, *then we have:*

(i) $L_{n,m}^c(e_0;x)=1$;

(ii) $L_{n,m}^c(e_1;x)=x$;

(iii)

$$L_{n,m}^c(e_2;x)=\frac{(n+(m+1)c)}{(n+(m-1)c)}x^2+\frac{2}{(n+(m-1)c)}x;$$

(iv)

$$L_{n,c}(e_3;x)=\frac{(n+(m+1)c)(n+(m+2)c)}{(n+(m-1)c)(n+(m-2)c)}x^3+\frac{6(n+(m+1)c)}{(n+(m-1)c)(n+(m-2)c)}x^2 + \frac{6}{(n+(m-1)c)(n+(m-2)c)}x;$$

(v)

$$L_{n,c}(e_4;x)=\frac{(n+(m+1)c)(n+(m+2)c)(n+(m+3)c)}{(n+(m-1)c)(n+(m-2)c)(n+(m-3)c)}x^4$$

$$+\frac{12(n+(m+1)c)(n+(m+2)c)}{(n+(m-1)c)(n+(m-2)c)(n+(m-3)c)}x^3$$
$$+\frac{36(n+(m+1)c)}{(n+(m-1)c)(n+(m-2)c)(n+(m-3)c)}x^2$$
$$+\frac{24}{(n+(m-1)c)(n+(m-2)c)(n+(m-3)c)}x.$$

All the moments of operators (1) can be obtained in terms of hyper geometric function of order $r \in \mathbb{N}$. For details see [7].

LEMMA 2. *The central moment of the operators (1) are given as*

$$\mu^c_{n,s}(x) = L_{n,c}((t-x)^s;x),$$

for $s = 2,4$, then we have:

(i)

$$\mu^c_{n,2}(x) = \frac{2x(1+cx)}{(n+(m-1)c)};$$

(ii)

$$\mu^c_{n,4}(x) = \frac{12c^2(n+(m+7)c)}{(n+(m-1)c)(n+(m-2)c)(n+(m-3)c)}x^4$$
$$+\frac{24c^2(13n+(13m+1)c)}{(n+(m-1)c)(n+(m-2)c)(n+(m-3)c)}x^3$$
$$+\frac{12(n+(m+9)c)}{(n+(m-1)c)(n+(m-2)c)(n+(m-3)c)}x^2$$
$$+\frac{24}{(n+(m-1)c)(n+(m-2)c)(n+(m-3)c)}x.$$

REMARK 1. If n is sufficiently large, then the central moment of the operators (1) are

$$\mu^c_{n,2}(x) \leqslant C\frac{x(1+cx)}{n}$$

and

$$\mu^c_{n,4}(x) \leqslant C\frac{(x(1+cx))^2}{n^2},$$

where $C > 0$ is constant.

REMARK 2. We know that $\sum_{j=0}^{\infty} p_{n+mc,j}(x,c) = 1$ and from (2), we have

$$F^{c,\alpha}_{n,m}(1;x) = \{n+(m+1)c\}\sum_{k=1}^{\infty} Q^{(\alpha)}_{n+mc}(x;c)\int_0^{\infty} p_{n+(m+2)c,k-1}(t;c)dt + Q^{(\alpha)}_{n+mc,0}(x;c)$$

$$= \sum_{k=0}^{\infty} Q_{n+mc}^{(\alpha)}(x;c) = \left(J_{n+mc,0}(x,c)\right)^{(\alpha)} = \left(\sum_{k=0}^{\infty} p_{n+mc,j}(x,c)\right)^{(\alpha)} = 1.$$

LEMMA 3. *For each* $f \in C_B[0,\infty)$ *then we have*

$$\left|F_{n,m}^{c,\alpha}(f(t);x)\right| \leqslant \|f\|.$$

Proof. It is easy to prove the above result by using Remark 2, therefore we skip the proof. □

LEMMA 4. *For every* $f \in C_B[0,\infty)$ *then we have*

$$\left|F_{n,m}^{c,\alpha}(f(t);x)\right| \leqslant \alpha L_{n,m}^{c}(\|f\|;x).$$

Proof. For $0 \leqslant c \leqslant d \leqslant 1$, $\alpha \geqslant 1$, using the inequality

$$|c^{\alpha} - d^{\alpha}| \leqslant \alpha |c-d|,$$

from the definition of $Q_{n+mc,k}^{(\alpha)}(x;c)$, for all $k \in \mathbb{N} \cup \{0\}$, we get

$$\begin{aligned} 0 &< \left(J_{n+mc,k}(x,c)\right)^{\alpha} - \left(J_{n+mc,k+1}(x,c)\right)^{\alpha} \leqslant \alpha\left(J_{n+mc,k}(x,c) - J_{n+mc,k+1}(x,c)\right) \\ &= \alpha p_{n+mc}(x,c). \end{aligned}$$

Hence

$$\left|F_{n,m}^{c,\alpha}(f(t);x)\right| \leqslant \alpha L_{n,m}^{c}(\|f\|;x). \quad \square$$

3. Main theorems

For $x > 0$, $t \geqslant 0$ and $0 < \gamma \leqslant 1$, we can see in Özarslan et al. [13], the Lipschitz type space is defined as:

$$Lip_M(\gamma) := \left\{ f \in C[0,\infty) : |f(t) - f(x)| \leqslant M \frac{|t-x|^{\gamma}}{(t+x)^{\gamma/2}} \right\}.$$

Now we estimate the rate of convergence of the function $f \in Lip_M(\gamma)$ by the operators $F_{n,m}^{c,\alpha}(.;x)$.

THEOREM 1. (Main) *For* $f \in Lip_M(\gamma)$ *and* $\gamma \in (0,1]$. *Then for* $x > 0$ *we have*

$$\left|F_{n,m}^{c,\alpha}(f(t);x) - f(x)\right| \leqslant \alpha M \left(\frac{\delta_{n,m}^{c}(x)}{x}\right)^{\gamma/2},$$

where $\delta_{n,m}^{c}(x) = \sqrt{\frac{2x(1+cx)}{(n+(m-1)c)}}$.

Proof. Using Lemma 4, we get

$$\begin{aligned}\left|F_{n,m}^{c,\alpha}(f(t);x)-f(x)\right| &\leqslant F_{n,m}^{c,\alpha}(|f(t)-f(x)|;x) \leqslant \alpha L_{n,m}^{c}(|f(t)-f(x)|;x)\\ &\leqslant \alpha M L_{n,m}^{c}\left(\frac{|t-x|^{\gamma}}{(t+x)^{\gamma/2}};x\right) \leqslant \frac{\alpha M}{x^{\gamma/2}} L_{n,m}^{c}\left(|t-x|^{\gamma};x\right). \end{aligned} \tag{4}$$

Using Hölder's inequality by taking $p=2/\gamma$ and $q=2/(2-\gamma)$, we get

$$\begin{aligned} L_{n,m}^{c}\left(|t-x|^{\gamma};x\right) &\leqslant \left\{L_{n,m}^{c}\left((t-x)^{2};x\right)\right\}^{\frac{\gamma}{2}}.\left\{L_{n,m}^{c}\left(1^{\frac{2}{(2-\gamma)}};x\right)\right\}^{\frac{(2-\gamma)}{2}}\\ &\leqslant \left\{L_{n,m}^{c}\left((t-x)^{2};x\right)\right\}^{\frac{\gamma}{2}} = \left(\delta_{n,m}^{c}(x)\right)^{\frac{\gamma}{2}}. \end{aligned} \tag{5}$$

From (4) and (5), we get

$$\left|F_{n,m}^{c,\alpha}(f(t);x)-f(x)\right| \leqslant \alpha M\left(\frac{\delta_{n,m}^{c}(x)}{x}\right)^{\frac{\gamma}{2}}.$$

Hence the proof. □

In our next theorem, we estimate the rate of convergence by using of Ditzian-Totik modulus of smoothness $\omega_{\phi^{\beta}}(f,\delta)$ and Peetre $K-$functional $K_{\phi^{\beta}}(f,\delta)$, $0\leqslant\beta\leqslant 1$. For $f\in C_B[0,\infty)$ and $\phi(x)=\sqrt{x(1+cx)}$, the Ditzian-Totik modulus of smoothness is defined as

$$\omega_{\phi^{\beta}}(f,\delta)=\sup_{0\leqslant i\leqslant\delta}\sup_{x\pm\frac{i\phi^{\beta}(x)}{2}\in[0,\infty)}\left|f\left(x+\frac{i\phi^{\beta}(x)}{2}\right)-f\left(x-\frac{i\phi^{\beta}(x)}{2}\right)\right|$$

and the Peetre $K-$functional is defined as

$$\omega_{\phi^{\beta}}(f,\delta)=\inf_{g\in W_{\beta}}\left\{\|f-g\|-\delta\left\|\phi^{\beta}g'\right\|\right\},$$

where W_{β} is subspace of the space which is locally absolutely continuous functions g on $[0,\infty)$, with the normed $\left\|\phi^{\beta}g'\right\|\leqslant\infty$. In [6, Theorem 2.1.1], there exists a constant $C>0$ such that

$$C^{-1}\omega_{\phi^{\beta}}(f,\delta)\leqslant K_{\phi^{\beta}}(f,\delta)\leqslant C\omega_{\phi^{\beta}}(f,\delta). \tag{6}$$

THEOREM 2. (Main) *For $f\in C_B[0,\infty)$ then we have*

$$\left|F_{n,m}^{c,\alpha}(f(t);x)-f(x)\right|\leqslant C\omega_{\phi^{\beta}}\left(f;\frac{\phi^{1-\beta}(x)}{\sqrt{n}}\right),$$

for sufficiently large n and C is a positive constant independent from f and n.

Proof. For $g \in W_\beta$, we get

$$g(t) = g(x) + \int_x^t g'(u)du. \tag{7}$$

Applying $F_{n,m}^{c,\alpha}$ in (7) and using Hölder's inequality then, we have

$$\begin{aligned}\left|F_{n,m}^{c,\alpha}(g(t);x) - g(x)\right| &\leqslant F_{n,m}^{c,\alpha}\left(\int_x^t \left|g'\right|du;x\right) \leqslant \left\|\phi^\beta g'\right\| F_{n,m}^{c,\alpha}\left(\left|\int_x^t \frac{du}{\phi^\beta(u)}\right|;x\right) \\ &\leqslant \left\|\phi^\beta g'\right\| F_{n,m}^{c,\alpha}\left(|t-x|^{1-\beta}\left|\int_x^t \frac{du}{\phi(u)}\right|^\beta;x\right).\end{aligned} \tag{8}$$

Let us take $A = \left|\int_x^t \frac{du}{\phi(u)}\right|$, then we get

$$\begin{aligned}A &\leqslant \left|\int_x^t \frac{du}{\sqrt{u}}\right|\left|\left(\frac{1}{\sqrt{1+cx}} + \frac{1}{\sqrt{1+ct}}\right)\right| \leqslant 2\left|\sqrt{t}-\sqrt{x}\right|\left(\frac{1}{\sqrt{1+cx}} + \frac{1}{\sqrt{1+ct}}\right) \\ &\leqslant 2\frac{|t-x|}{\sqrt{x}+\sqrt{t}}\left(\frac{1}{\sqrt{1+cx}} + \frac{1}{\sqrt{1+ct}}\right) \leqslant 2\frac{|t-x|}{\sqrt{x}}\left(\frac{1}{\sqrt{1+cx}} + \frac{1}{\sqrt{1+ct}}\right).\end{aligned} \tag{9}$$

The inequality $|a+b|^\beta \leqslant |a|^\beta + |b|^\beta$, $0 \leqslant \beta \leqslant 1$, then from (9) we get

$$\left|\int_x^t \frac{du}{\phi(u)}\right|^\beta \leqslant 2^\beta \frac{|t-x|^\beta}{x^{\beta/2}}\left(\frac{1}{(1+cx)^{\beta/2}} + \frac{1}{(1+ct)^{\beta/2}}\right). \tag{10}$$

From (8), (10) and using Cauchy inequality then we get

$$\begin{aligned}\left|F_{n,m}^{c,\alpha}(g(t);x) - g(x)\right| &\leqslant \frac{2^\beta \left\|\phi^\beta g'\right\|}{x^{\beta/2}} F_{n,m}^{c,\alpha}\left(|t-x|\left(\frac{1}{(1+cx)^{\beta/2}} + \frac{1}{(1+ct)^{\beta/2}}\right);x\right) \\ &\leqslant \frac{2^\beta \left\|\phi^\beta g'\right\|}{x^{\beta/2}}\left(\frac{1}{(1+cx)^{\beta/2}}\left(F_{n,c}^{(\alpha)}\left((t-x)^2;x\right)\right)^{1/2}\right. \\ &\quad \left. + \left(F_{n,m}^{c,\alpha}\left((t-x)^2;x\right)\right)^{1/2}.\left(F_{n,m}^{c,\alpha}\left((1+ct)^{-\beta};x\right)\right)^{1/2}\right).\end{aligned} \tag{11}$$

If n is sufficiently large then we get

$$\left(F_{n,m}^{c,\alpha}\left((t-x)^2;x\right)\right)^{1/2} \leqslant \sqrt{\frac{2\alpha}{n}}\phi(x), \tag{12}$$

where $\phi(x)=\sqrt{x(1+cx)}$. For each $x\in[0,\infty)$, $F_{n,m}^{c,\alpha}\left((1+ct)^{-\beta};x\right)\to(1+cx)^{-\beta}$ as $n\to\infty$. Thus for $\varepsilon>0$, there exists $n_0\in\mathbb{N}$ such that

$$F_{n,m}^{c,\alpha}\left((1+ct)^{-\beta};x\right)\leqslant(1+cx)^{-\beta}+\varepsilon,\quad \text{for all } n\geqslant n_0.$$

By choosing $\varepsilon=(1+cx)^{-\beta}$ then, we get

$$F_{n,m}^{c,\alpha}\left((1+ct)^{-\beta};x\right)\leqslant 2(1+cx)^{-\beta},\quad \text{for all } n\geqslant n_0. \tag{13}$$

From (11) to (13), we have

$$\begin{aligned}\left|F_{n,m}^{c,\alpha}(g(t);x)-g(x)\right|&\leqslant 2^{\beta}\left\|\phi^{\beta}g'\right\|\sqrt{\frac{2\alpha}{n}}\phi(x)\left(\phi^{-\beta}(x)+\sqrt{2}x^{-\frac{\beta}{2}}(1+cx)^{-\frac{\beta}{2}}\right)\\&\leqslant 2^{\beta+\frac{1}{2}}(1+\sqrt{2})\left\|\phi^{\beta}g'\right\|\sqrt{\frac{\alpha}{n}}\phi^{1-\beta}(x).\end{aligned} \tag{14}$$

We may write

$$\begin{aligned}\left|F_{n,m}^{c,\alpha}(f(t);x)-f(x)\right|&\leqslant\left|F_{n,m}^{c,\alpha}(f(t)-g(t);x)\right|+\left|F_{n,m}^{c,\alpha}(g(t);x)-g(x)\right|+|g(x)-f(x)|\\&\leqslant 2\|f-g\|+\left|F_{n,m}^{c,\alpha}(g(t);x)-g(x)\right|.\end{aligned} \tag{15}$$

From (14) to (15) and for sufficiently large n, we get

$$\begin{aligned}\left|F_{n,m}^{c,\alpha}(f(t);x)-f(x)\right|&\leqslant 2\|f-g\|+2^{\beta+\frac{1}{2}}(1+\sqrt{2})\sqrt{\frac{\alpha}{n}}\phi^{1-\beta}(x)\left\|\phi^{\beta}g'\right\|\\&\leqslant C_1\left\{\|f-g\|+\frac{\phi^{1-\beta}(x)}{\sqrt{n}}\left\|\phi^{\beta}g'\right\|\right\}\leqslant CK_{\phi^{\beta}}\left(f,\frac{\phi^{1-\beta}(x)}{\sqrt{n}}\right),\end{aligned} \tag{16}$$

where $C_1=max(2,2^{\beta+\frac{1}{2}}(1+\sqrt{2})\sqrt{\alpha})$ and $C=2C_1$. From (6) and (16), we get the required result. □

For the estimation of the rate of convergence of the function $f\in C_2[0,\infty)$ by using the weighted modulus of continuity, which was introduced by Ispir and Yüksal [18] as follows:

$$\Omega(f;\delta)=\sup_{x\in[0,\infty),0<\beta<\delta}\frac{f(x+\beta)-f(x)}{1+(x+\beta)^2}. \tag{17}$$

Many authors have discussed weighted modulus of continuity for various linear positive operators. For more information (see [2], [3], [5]).

There are several properties of weighted modulus of continuity $\Omega(.;\delta)$ which are stated in following Lemma.

LEMMA 5. [18] *For* $f \in C_2[0,\infty)$ *the following properties hold:*

(i) $\Omega(f;\delta)$ *is monotonically increasing in* δ;

(ii) $\lim_{\delta\to 0^+} \Omega(f;\delta) = 0$;

(iii) for each $r \in \mathbb{N}$, $\Omega(f;r\delta) \leqslant r\Omega(f;\delta)$;

(iv) for each $\lambda \in [0,\infty)$, $\Omega(f;\lambda\delta) \leqslant (\lambda+1)\Omega(f;\delta)$.

THEOREM 3. (Main) *Let* $f \in C_2[0,\infty)$, $\alpha > 0$, *for fixed* m *and sufficiently large* n *then we have*

$$\sup_{x\in[0,\infty)} \frac{\left|F_{n,m}^{c,\alpha}(f;x) - f(x)\right|}{(1+x)^{5/2}} \leqslant C\Omega\left(f;\frac{1}{\sqrt{n}}\right),$$

where C *is positive constant depends on* n *and* f.

Proof. By the definition of the weighted modulus of continuity and Lemma 5, we have

$$\begin{aligned} |f(t)-f(x)| &\leqslant \left(1+(x+|t-x|)^2\right)\Omega\left(f;|t-x|\right) \\ &\leqslant 2(1+x^2)\left(1+(t-x)^2\right)\left(1+\frac{|t-x|}{\delta}\right)\Omega(f;\delta). \end{aligned} \tag{18}$$

Applying $F_{n,m}^{c,\alpha}(.;x)$ on both side of (18), we can write

$$\left|F_{n,m}^{c,\alpha}(f;x) - f(x)\right| \leqslant \left[1 + F_{n,m}^{c,\alpha}((t-x)^2;x) + F_{n,m}^{c,\alpha}\left((1+(t-x)^2)\frac{|t-x|}{\delta};x\right)\right]. \tag{19}$$

From Remark 1 and using Cauchy-Schwarz inequality in the last term of (19), we have

$$F_{n,m}^{c,\alpha}\left((1+(t-x)^2)\frac{|t-x|}{\delta};x\right) \leqslant \frac{1}{\delta}\left(\alpha\mu_{n,2}^{c}(x)\right)^{1/2} + \frac{1}{\delta}\left(\alpha\mu_{n,4}^{c}(x)\right)^{1/2}\left(\alpha\mu_{n,2}^{c}(x)\right)^{1/2}. \tag{20}$$

Combining the estimate from (18) to (20) and taking $C = 2(1+\sqrt{\alpha C}+2C)$ and $\delta = \frac{1}{\sqrt{n}}$ then we get the required result. □

In next theorem, we study the rate of convergence of the Bézier variant of Gupta-Srivastava operators (2) in the class $DBV[0,\infty)$, the class of all absolutely continuous functions f defined on $[0,\infty)$ having a derivative coinciding a.e. with a function of bounded variation on $[0,\infty)$. It can be seen that for $f \in DBV[0,\infty)$, we can write

$$f(x) = \int_0^x g(t)\,dt + f(0),$$

where $g(t)$ is a function of bounded variation on each finite subinterval of $[0,\infty)$. The operators (2) can be rewritten in the following form:

$$F_{n,m}^{c,\alpha}(f(t);x) = \int_0^\infty M_{n,m,c}^{(\alpha)}(x,t) f(t)dt, \tag{21}$$

where

$$M_{n,m,c}^{(\alpha)}(x,t) = \{n+(m+1)c\} \sum_{k=1}^{\infty} Q_{n+mc,k}^{(\alpha)}(x;c) p_{n+(m+2)c,k}(t,c) + Q_{n+mc,0}^{(\alpha)}(x;c)\delta(t),$$

where $\delta(t)$ is Dirac delta function.

LEMMA 6. *For a fixed $x \in [0,\infty)$ and n sufficiently large then we have:*

(i) $\zeta_{n,c}^{(\alpha)}(x;y) = \int_0^y M_{n,m,c}^{(\alpha)}(x;t)dt \leqslant \frac{2\alpha x(1+cx)}{n(x-y)^2}, \quad 0 \leqslant y \leqslant x;$

(ii) $1-\zeta_{n,c}^{(\alpha)}(x;z) = \int_z^\infty M_{n,m,c}^{(\alpha)}(x;t)dt \leqslant \frac{2\alpha x(1+cx)}{n(z-x)^2}, \quad x \leqslant z \leqslant \infty.$

Proof. From (21) and using Remark 1 then we have

$$\zeta_{n,c}^{(\alpha)}(x;y) \leqslant \int_0^y M_{n,m,c}^{(\alpha)}(x;t)\left(\frac{x-t}{x-y}\right)^2 dt \leqslant \frac{\alpha}{(x-y)^2} L_{n,c}\left((e_1-x)^2;x\right) \leqslant \frac{2\alpha x(1+cx)}{n(x-y)^2}.$$

We can prove the second part of Lemma in same way. □

THEOREM 4. *Let $f \in DBV[0,\infty)$ then for every $x \in [0,\infty)$ and n sufficiently large we have*

$$\begin{aligned}
\left|F_{n,m}^{c,\alpha}(f;x) - f(x)\right| \leqslant & \frac{1}{\alpha+1}\left|f'(x+) + \alpha f'(x-)\right| \sqrt{\frac{2\alpha x(1+cx)}{n}} \\
& + \frac{\alpha}{\alpha+1}\left|f'(x+) - f'(x-)\right| \sqrt{\frac{2\alpha x(1+cx)}{n}} \\
& + \frac{2\alpha(1+cx)}{n} \sum_{k=1}^{[\sqrt{n}]} \bigvee_{x-\frac{x}{k}}^{x} f_x' + \frac{x}{\sqrt{n}} \bigvee_{x-\frac{x}{\sqrt{n}}}^{x} f_x' \\
& + \frac{2\alpha(1+cx)}{nx}\left|f(2x) - f(x) - xf(x+)\right| \\
& + \frac{2\alpha x(1+cx)}{n} \sum_{k=1}^{[\sqrt{n}]} \bigvee_{x}^{x+\frac{x}{k}} f_x' + \frac{x}{\sqrt{n}} \bigvee_{x}^{x+\frac{x}{\sqrt{n}}} (f_x') + M(\gamma,r,x) \\
& + \frac{2\alpha(1+cx)}{nx}\left|f(x)\right| + \sqrt{\frac{2\alpha x(1+cx)}{n}}\left|f(x+)\right|,
\end{aligned}$$

where $\bigvee_a^b f(x)$ *denotes the total variation of* f *on* $[a,b]$, f_x *is an auxiliary operator given by*

$$f_x(t) = \begin{cases} f(t) - f(x-), & 0 \leqslant t < x, \\ 0, & t = x, \\ f(t) - f(x+), & x < t < \infty. \end{cases} \tag{22}$$

Proof. From Remark 2, $F_{n,m}^{c,\alpha}(1;x) = 1$ and using the alternative form of the operators (21) for each $x \in [0,\infty)$ then we have

$$F_{n,m}^{c,\alpha}(f(t);x) - f(x) = \int_0^\infty M_{n,m,c}^{(\alpha)}(x,t)\big(f(t) - f(x)\big)dt = \int_0^\infty M_{n,m,c}^{(\alpha)}(x,t)\left(\int_x^t f'(u)du\right)dt. \tag{23}$$

For each $f \in DBV[0,\infty)$ and from (22), we can write

$$\begin{aligned} f'(u) =& f_x'(u) + \frac{1}{\alpha+1}(f'(x+) + \alpha f'(x-)) \\ &+ \frac{1}{2}\left(f'(x+) + \alpha f'(x-)\right)\left(\operatorname{sgn}(u-x) + \frac{\alpha-1}{\alpha+1}\right) \\ &\times \delta_x(u)\left[f'(u) - \left(f'(x+) + f'(x-)\right)\right], \end{aligned} \tag{24}$$

where

$$\delta_x(u) = \begin{cases} 1, & u = x, \\ 0, & u \neq x. \end{cases}$$

From (23) and (24) we have

$$\begin{aligned} F_{n,m}^{c,\alpha}(f(t);x) - f(x) = \int_0^\infty M_{n,m,c}^{(\alpha)}(x,t) \int_x^t \Bigg(& f_x'(u) + \frac{1}{\alpha+1}(f'(x+) + \alpha f'(x-)) \\ &+ \frac{1}{2}(f'(x+) + \alpha f'(x-))\left(\operatorname{sgn}(u-x) + \frac{\alpha-1}{\alpha+1}\right) \\ &\times \delta_x(u)[f'(u) - \frac{1}{2}(f'(x+) + f'(x-)]\Bigg) dudt. \end{aligned} \tag{25}$$

It is easy to say that

$$\int_0^\infty M_{n,m,c}^{(\alpha)}(x,t) \int_x^t [f'(u) - \frac{1}{2}(f'(x+) + f'(x-)]\delta_x(u)dudt = 0. \tag{26}$$

Now

$$\begin{aligned} B_1 &= \int_0^{\infty} M_{n,m,c}^{(\alpha)}(x,t) \int_x^t \frac{1}{\alpha+1}(f'(x+)+\alpha f'(x-))dudt \\ &= \frac{1}{\alpha+1}(f'(x+)+\alpha f'(x-)) \int_0^{\infty} M_{n,m,c}^{(\alpha)}(x,t)(t-x)dt \\ &= \frac{1}{\alpha+1}(f'(x+)+\alpha f'(x-))F_{n,c}^{(\alpha)}((t-x);x) \end{aligned} \tag{27}$$

and

$$\begin{aligned} B_2 &= \int_0^{\infty} M_{n,m,c}^{(\alpha)}(x,t) \int_x^t \frac{1}{2}(f'(x+)+\alpha f'(x-)) \left(\operatorname{sgn}(u-x)+\frac{\alpha-1}{\alpha+1}\right)dudt \\ &= \frac{1}{2}(f'(x+)+\alpha f'(x-)) \left(-\int_0^x M_{n,m,c}^{(\alpha)}(x,t) \int_x^t \left(\operatorname{sgn}(u-x)+\frac{\alpha-1}{\alpha+1}\right)dudt \right. \\ &\quad \left. + \int_x^{\infty} M_{n,m,c}^{(\alpha)}(x,t) \int_x^t \left(\operatorname{sgn}(u-x)+\frac{\alpha-1}{\alpha+1}\right) \right) \\ &\leqslant \frac{\alpha}{\alpha+1}(f'(x+)+\alpha f'(x-)) \int_0^{\infty} M_{n,m,c}^{(\alpha)}(x,t)\,|t-x|\,dt \\ &\leqslant \frac{\alpha}{\alpha+1}(f'(x+)+\alpha f'(x-)) \left(F_{n,c}^{(\alpha)}\left((e_1-x)^2;x\right)\right)^{\frac{1}{2}}. \end{aligned} \tag{28}$$

By using Remark 1 and Lemma 4, from (25) – (28) then we have

$$\begin{aligned} F_{n,m}^{c,\alpha}(f;x) - f(x) \leqslant &\left|A_n^{(\alpha)}(f';x)+B_n^{(\alpha)}(f';x)\right| + \frac{2\alpha}{\alpha+1}\left|f'(x+)+\alpha f'(x-)\right| \frac{x(1+cx)}{n} \\ &+ \frac{\alpha}{\alpha+1}\left|f'(x+)-f'(x-)\right| \sqrt{\frac{2\alpha x(1+cx)}{n}}, \end{aligned} \tag{29}$$

where

$$A_n^{(\alpha)}(f';x) = \int_0^x \left(\int_x^t f_x'(u)du \right) M_{n,m,c}^{(\alpha)}(x,t)dt$$

and

$$B_n^{(\alpha)}(f';x) = \int_x^{\infty} \left(\int_x^t f_x'(u)du \right) M_{n,m,c}^{(\alpha)}(x,t)dt.$$

To estimate $A_n^{(\alpha)}(f';x)$, using integration by parts and applying Lemma 6 with $y = x - \frac{x}{\sqrt{n}}$, we obtain

$$\begin{aligned} A_n^{(\alpha)}(f';x) &= \left| \int_0^x \left(\int_x^t f_x'(u)du \right) d_t\, \zeta_{n,c}^{(\alpha)}(x;t) \right| = \left| \int_0^x \zeta_{n,c}^{(\alpha)}(x;t) f_x'(t)dt \right| \\ &\leqslant \int_0^y \left| f_x'(t) \right| \left| \zeta_{n,c}^{(\alpha)}(x;t) \right| dt + \int_0^y \left| f_x'(t) \right| \left| \zeta_{n,c}^{(\alpha)}(x;t) \right| dt \\ &\leqslant \frac{2\alpha x(1+cx)}{n} \int_0^y \overset{x}{\underset{t}{V}} f_x'(x-t)^2 dt + \int_y^x \overset{x}{\underset{t}{V}} f_x' dt \\ &\leqslant \frac{2\alpha x(1+cx)}{n} \int_0^{x-\frac{x}{\sqrt{n}}} \overset{x}{\underset{t}{V}} f_x'(x-t)^2 dt + \frac{x}{\sqrt{n}} \overset{x}{\underset{x-\frac{x}{\sqrt{n}}}{V}} f_x'. \end{aligned} \tag{30}$$

Substituting $u = \frac{x}{x-t}$, we get

$$\begin{aligned} \frac{2\alpha x(1+cx)}{n} \int_0^{x-\frac{x}{\sqrt{n}}} \overset{x}{\underset{t}{V}} f_x'(x-t)^2 dt &= \frac{2\alpha x(1+cx)}{nx} \int_1^{\sqrt{n}} \overset{x}{\underset{x-\frac{x}{u}}{V}} f_x' du \\ &\leqslant \frac{2\alpha(1+cx)}{n} \sum_{k=1}^{[\sqrt{n}]} \int_k^{k+1} \overset{x}{\underset{x-\frac{x}{k}}{V}} f_x' du \\ &\leqslant \frac{2\alpha(1+cx)}{n} \sum_{k=1}^{[\sqrt{n}]} \overset{x}{\underset{x-\frac{x}{k}}{V}} f_x'. \end{aligned} \tag{31}$$

From (30) and (31) we get

$$A_n^{(\alpha)}(f';x) = \frac{2\alpha(1+cx)}{n} \sum_{k=1}^{[\sqrt{n}]} \overset{x}{\underset{x-\frac{x}{k}}{V}} f_x' + \frac{x}{\sqrt{n}} \overset{x}{\underset{x-\frac{x}{\sqrt{n}}}{V}} f_x'. \tag{32}$$

We can write

$$B_n^{(\alpha)}(f';x) \leqslant \left| \int_x^{2x} \left(\int_x^t f_x'(u)du \right) d_t(1 - \zeta_{n,c}^{(\alpha)}(x;t)) \right| + \left| \int_{2x}^{\infty} \left(\int_x^t f_x'(u)du \right) d_t M_{n,m,c}^{(\alpha)}(x,t) \right|.$$

From the second part of the Lemma 6, we get

$$M_{n,m,c}^{(\alpha)}(x,t) = d_t((1 - \zeta_{n,c}^{(\alpha)}(x;t)), \qquad \text{for } t > x.$$

Hence

$$B_n^{(\alpha)}(f';x) = B_{n,1}^{(\alpha)}(f';x) + B_{n,2}^{(\alpha)}(f';x),$$

where

$$B_{n,1}^{(\alpha)}(f';x)=\left|\int_x^{2x}\left(\int_x^t f_x'(u)du\right)d_t(1-\zeta_{n,c}^{(\alpha)}(x;t))\right|$$

and

$$B_{n,2}^{(\alpha)}(f';x)=\left|\int_{2x}^{\infty}\left(\int_x^t f_x'(u)du\right)d_tM_{n,m,c}^{(\alpha)}(x,t)\right|.$$

Using integration by parts, applying Lemma 6, $1-\zeta_{n,c}^{(\alpha)}(x;t)\leqslant 1$ and taking $t=x+\frac{x}{u}$ successively,

$$\begin{aligned}
B_{n,1}^{(\alpha)}(f';x)&=\left|\int_x^{2x}f_x'(u)du(1-\zeta_{n,c}^{(\alpha)}(x;2x))-\int_x^{2x}f_x'(t)(1-\zeta_{n,c}^{(\alpha)}(x;t))dt\right|\\
&\leqslant\left|\int_x^{2x}(f'(u)-f'(x+))du\right|\left|1-\zeta_{n,c}^{(\alpha)}(x;2x)\right|+\left|\int_x^{2x}f_x'(t)(1-\zeta_{n,c}^{(\alpha)}(x;t))dt\right|\\
&\leqslant\frac{2\alpha(1+cx)}{nx}\left|f(2x)-f(x)-xf(x+)\right|\\
&\quad+\frac{2\alpha x(1+cx)}{n}\int_{x+\frac{x}{\sqrt{n}}}^{2x}\frac{\overset{t}{V}_x f_x'}{(t-x)^2}dt+\int_x^{x+\frac{x}{\sqrt{n}}}\overset{t}{V}_x f_x'dt\\
&\leqslant\frac{2\alpha(1+cx)}{nx}\left|f(2x)-f(x)-xf(x+)\right|\\
&\quad+\frac{2\alpha x(1+cx)}{n}\sum_{k=1}^{[\sqrt{n}]}\overset{x+\frac{x}{k}}{V}_x f_x'+\frac{x}{\sqrt{n}}\overset{x+\frac{x}{\sqrt{n}}}{V}_x(f_x').
\end{aligned}\tag{33}$$

Using Remark 1 then we have

$$\begin{aligned}
B_{n,2}^{(\alpha)}(f';x)&=\left|\int_{2x}^{\infty}\left(\int_x^t(f'(u)-f'(x+))du\right)M_{n,m,c}^{(\alpha)}(x,t)dt\right|\\
&\leqslant\int_{2x}^{\infty}|f(t)-f(x)|M_{n,m,c}^{(\alpha)}(x,t)dt+\int_{2x}^{\infty}|t-x|\,|f(x+)|\,M_{n,m,c}^{(\alpha)}(x,t)dt\\
&\leqslant\left|\int_{2x}^{\infty}f(t)M_{n,m,c}^{(\alpha)}(x,t)dt\right|+|f(x)|\left|\int_{2x}^{\infty}M_{n,m,c}^{(\alpha)}(x,t)dt\right|\\
&\quad+|f(x+)|\left(\int_{2x}^{\infty}(e_1-x)^2M_{n,m,c}^{(\alpha)}(x,t)dt\right)^{\frac{1}{2}}
\end{aligned}\tag{34}$$

$$\leqslant M\int_{2x}^{\infty} t^{\gamma} M_{n,m,c}^{(\alpha)}(x,t)dt + |f(x)|\left|\int_{2x}^{\infty} M_{n,m,c}^{(\alpha)}(x,t)dt\right| + \sqrt{\frac{2\alpha x(1+cx)}{n}}\,|f(x+)|\,.$$

For $t \geqslant 2x$, we get $t \leqslant 2(t-x)$ and $x \leqslant t-x$, applying Hölder's inequality, we have

$$\begin{aligned} B_{n,2}^{(\alpha)}(f';x) \leqslant & M2^{\gamma}\left(\int_{2x}^{\infty}(e_1-x)^{2r}M_{n,m,c}^{(\alpha)}(x,t)dt\right)^{\frac{\gamma}{2r}} \\ & + \frac{2\alpha(1+cx)}{nx}|f(x)| + \sqrt{\frac{2\alpha x(1+cx)}{n}}\,|f(x+)| \\ = & M(\gamma,c,r,x) + \frac{2\alpha(1+cx)}{nx}|f(x)| + \sqrt{\frac{2\alpha x(1+cx)}{n}}\,|f(x+)|\,. \end{aligned} \tag{35}$$

From (33) and (35), we get

$$\begin{aligned} B_n^{(\alpha)}(f';x) = & \frac{2\alpha(1+cx)}{nx}|f(2x)-f(x)-xf(x+)| \\ & + \frac{2\alpha x(1+cx)}{n}\sum_{k=1}^{[\sqrt{n}]}\overset{x+\frac{x}{k}}{\underset{x}{V}} f'_x + \frac{x}{\sqrt{n}}\overset{x+\frac{x}{\sqrt{n}}}{\underset{x}{V}}(f'_x) \\ & + M(\gamma,c,r,x) + \frac{2\alpha(1+cx)}{nx}|f(x)| + \sqrt{\frac{2\alpha x(1+cx)}{n}}\,|f(x+)|\,. \end{aligned} \tag{36}$$

From (29), (32) and (36), we get our desired result. □

4. Graphical results and discussion

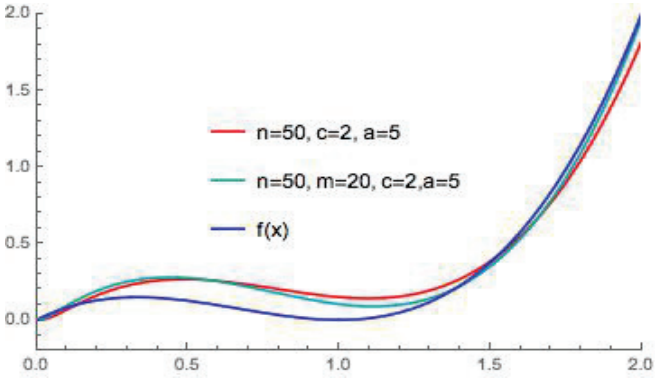


Figure 1: Comparison between Bézier variant of Srivastava-Gupta operators [8] (red) with operators (2)(cyan) along with function $f(x)$.

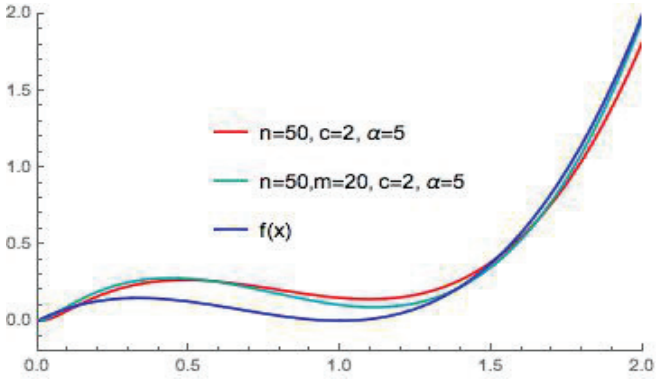


Figure 2: Comparison between Bézier variant of Srivastava-Gupta Operators [9] (red) with operators (2) (cyan) along with function $f(x)$ (blue).

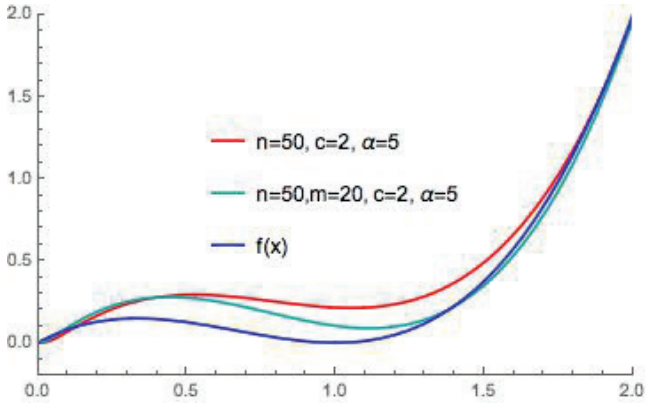


Figure 3: Comparison between Bézier variant of modified Srivastava-Gupta operators [12] (red) with operators (2) (cyan) along with function $f(x)$ (blue).

In operators (2) by taking $m=-1$, we obtain Bézier variant of Srivastava-Gupta operators which were propoesd by Ispir and Yüksel [8] and its modification by Neer et al.[12]. For $m=1$ in (2), we obtain another Bézier form of Srivastava-Gupta operators considered by Kajla [9]. The proposed operators (2) have generalized form with different values of m. Here, we show graphical comparison between operators (2) for $m=20$ with above discussed operators ([8], [9], [12]) for the function $f(x)=x^3-2x^2+x$.

From the above graphs we observe here that we have better approximation for the Bézier variant of Gupta-Srivastava operators (2), discussed in the present paper than the other variants of [15], therefore it is justified to study this form of operators.

Acknowledgement. The authors are thankful to the anonymous referees for their valuable comments and suggestions, leading to overall improvements in the paper.

REFERENCES

[1] T. ACAR, L. N. MISHRA AND V. N. MISHRA, *Simultaneous approximation for generalized Srivastava-Gupta operators*, J. Funct. Spaces, (2015), Article ID 936308, 1-11.

[2] P. N. AGRAWAL,S. ARACI, M. BOHNER AND L. KUMARI, *Approximation degree of Durrmeyer-Bézier type operators*, J. Ineql. Appl., (2018), 1-17.

[3] N. DEO, *A note on equivalent theorem for Beta operators*, Mediterr. j. math., **4**, (2007), 245-250.

[4] N. DEO, *Faster rate of convergence on Srivastava-Gupta operators*, Appl. Math. Comput., **21**, 218(2012), 10486-10491.

[5] M. DHAMIJA, R. PRATAP AND N. DEO, *Approximation by Kantorovich form modified Szász-Mirakyan Operators*. Appl Math Comput,**317**, (2018), 109-120.

[6] Z. DITZIAN AND V. TOTIK, *Moduli of smoothness*, Springer series in computational Mathematics, **9**, (1987), Springer New York.

[7] V. GUPTA AND H. M. SRIVASTAVA, *A General family of the Srivastava-Gupta operators preserving linear functions*, European J. Pure Appl. maths, **3**, 11(2018), 575-579.

[8] N. ISPIR AND I. YÜKSEL, *On the Bézier variant of Srivastava-Gupta operators*, Appl. Math. E-Notes, **5**, (2005), 129-137.

[9] A. KAJLA, *On the Bézier variant of the Srivastava-Gupta Operators*, Constructive Mathematical Analysis, **1**, (2018), 99-107.

[10] A. KUMAR, *Approximation by Stancu type generalized Srivastava-Gupta operators based on certain parameter*, Khayaam J. Math., **3**, (2017), 147-159.

[11] P. MAHESWARI (SHARMA), *On modified Srivastava-Gupta operators*, Filomat, **6**, 29(2015), 1173-1177.

[12] T. NEER, N. ISPIR AND P. N. AGRAWAL, *Bézier variant of modified Srivastava-Gupta operators*, Revista de la Unio In Matema Itica Argentina, **2**, 58(2017), 199-214.

[13] Özarslan M.A. and Duman O., *Local approximation behaviour of modified SMK operators*, Miskolk Math.Notes, **11(1)**(2010), 87-99.

[14] R. PRATAP AND N. DEO, *Approximation by Genuine Gupta-Srivastava Operators*, RACSAM, (2019), https://doi.org/10.1007/s13398-019-00633-4.

[15] H. M. SRIVASTAVA AND V. GUPTA, *A Certain family of summation-integral type operators*, Math. Comput. Modelling, **37**, (2003), 1307-1315.

[16] D. K. VERMA AND P. N. AGRAWAL, *Convergence in simultaneous approximation for Srivastava-Gupta operators*, Math. Sci., Springer, **6**, (2012), Article ID 22, 8p.

[17] R. YADAV, *Approximation by modified Srivastava-Gupta operators*, Appl. Math. Comput., **226**, (2014), 61-66.

[18] I. YÜKSEL AND N. ISPIR, *Weighted Approximation by a certain family of Summation integral-type operators*, Comput. Math. Appl., **52**, (10-11)(2007), 1463-1470.

(Received March 3, 2019)

Ram Pratap
Delhi Technological University, Formerly Delhi College of Engineering
Department of Applied Mathematics
Bawana Road, 110042 Delhi, India
e-mail: rampratapiitr@gmail.com

Naokant Deo
Delhi Technological University, Formerly Delhi College of Engineering
Department of Applied Mathematics
Bawana Road, 110042 Delhi, India
e-mail: naokantdeo@dce.ac.in

Artificial Intelligence Review
https://doi.org/10.1007/s10462-019-09794-5

Sentiment analysis using deep learning architectures: a review

Ashima Yadav[1] · **Dinesh Kumar Vishwakarma**[1]

© Springer Nature B.V. 2019

Abstract
Social media is a powerful source of communication among people to share their sentiments in the form of opinions and views about any topic or article, which results in an enormous amount of unstructured information. Business organizations need to process and study these sentiments to investigate data and to gain business insights. Hence, to analyze these sentiments, various machine learning, and natural language processing-based approaches have been used in the past. However, deep learning-based methods are becoming very popular due to their high performance in recent times. This paper provides a detailed survey of popular deep learning models that are increasingly applied in sentiment analysis. We present a taxonomy of sentiment analysis and discuss the implications of popular deep learning architectures. The key contributions of various researchers are highlighted with the prime focus on deep learning approaches. The crucial sentiment analysis tasks are presented, and multiple languages are identified on which sentiment analysis is done. The survey also summarizes the popular datasets, key features of the datasets, deep learning model applied on them, accuracy obtained from them, and the comparison of various deep learning models. The primary purpose of this survey is to highlight the power of deep learning architectures for solving sentiment analysis problems.

Keywords Datasets · Deep learning · Opinion mining · Review analysis · Sentiment analysis

1 Introduction

The Internet is serving as a universal and most cost-effective source of information complemented by the growth of social media. Blogs, reviews, tweets, posts, discussions on social media are scanned for extracting the opinion of people. The attitude, views, feelings, opinions constitute an essential part in analyzing the behavior of a person, which can be referred

✉ Dinesh Kumar Vishwakarma
dvishwakarma@gmail.com

Ashima Yadav
ashimayadavdtu@gmail.com

[1] Biometric Research Laboratory, Department of Information Technology, Delhi Technological University, New Delhi 110042, India

Published online: 02 December 2019

Table 1 Applications of sentiment analysis

Applications	Refs.
Financial market prediction	Bhardwaj et al. (2015), Napitu et al. (2017), Rao and Srivastava (2012), Xu and Kešelj (2014) and Bollen et al. (2011)
Business review analysis	Zvarevashe and Olugbara (2018), Singla et al. (2017), Hedge and Padma (2017), Haque et al. (2018), Xiong et al. (2018a) and Mataoui et al. (2018)
Politics	Haselmayer and Jenny (2017), Kušen and Strembeck (2017), Kumari and Babu (2017) and Wang et al. (2012a)
Demonetization	Singh et al. (2017), Arun et al. (2017), Roy et al. (2017) and Singh et al. (2018)
Crime prediction	Wang et al. (2012b), Rosenfeld and Fornango (2008), Gerber (2014) and Azeez and Aravindhar (2015)
Disaster assessment, response, management	Ragini et al. (2018), Beigi et al. (2016), Radianti et al. (2016) and Wu and Cui (2018)

as the *sentiments*. Sentiment analysis, also known as opinion mining, deals with inspecting these sentiments directed towards any entity. Liu (2010) used the term *object* to represent the target entity mentioned in the text. An object is constituted of *components* and some set of *attributes*. For example, consider the statement "*the screen of the laptop is damaged, and it has a terrible battery life*". The object here is a laptop having *screen* and *battery* as the components. *Display quality* is an attribute of the screen, and *battery life* is the attribute of the battery. The sentiments or opinions expressed in the text are further categorized into positive, negative, neutral, or into fine-grained classification (most positive, least positive, most negative, and least negative). Hence, in the above example, negative sentiment is conveyed for the laptop object. Moreover, sentiments can be expressed in the form of text, videos, audios, emoticons, and images.

1.1 Applications of sentiment analysis

Recently, it has been observed that the number of people actively involved in social media is rapidly increasing (Facebook Statistics 2019; Twitter Statistics 2019). People are expressing their opinions in the form of reviews, comments, posts, status on various topics. As a result, a tremendous amount of data is generated on the Internet, which can be analyzed for further research. This makes sentiment analysis a popular field having ample applications. Hence, to highlight the motivation behind the sentiment analysis and to create a more profound interest of users in this area of research, we briefly discuss the famous works in the field of sentiment analysis. Table 1 lists the widespread applications of sentiment analysis.

Stock market investment constitutes a dominant part of the economy of any country. Hence, a detailed market analysis becomes a crucial part of investing money in stock market. In Bhardwaj et al. (2015), a system is developed, which fetches the live data values of Sensex and Nifty, that serves as an essential indicators of stock market. Pre-processing and feature selection are also applied to the data followed by the sentiment analysis task to get stock market status. Rao and Srivastava (2012) analyzed the relationship between the sentiment of tweets from DJIA (Dow Jones Industrial Average), NASDAQ-100, and 13 other companies and its impact on market performance. They have used Naïve Bayesian classifier for sentiment

classification. The results show that the polarity of sentiments has a substantial impact on the stock price movement. Moreover, the sentiments of the previous week strongly impact the coming week's opening and closing stock values. Xu and Kešelj (2014) proposed a method in which SVM was used for the two-stage sentiment classification process, and the dataset consisted of Tweets from the StockTwits website. In the first stage, neutral and polarized classification was performed, followed by binary (positive and negative) classification. Then a schema was designed to analyze the labeled tweets. The experimental results showed that the overnight activity of the user on StockTwits had a positive correlation to the stock trading volume of the next day. Also, it was found that the collective sentiment of after-hours (4:00 pm–9:30 am) influenced the stock movement direction of the following day.

A significant goal of sentiment analysis is to classify and analyze the reviews related to products, hotels, online booking sites, e-commerce sites, social media, etc. Haque et al. (2018) used Amazon product reviews in three domains: '*cell phone and accessories*', '*musical*', and '*electronics*'. They have classified the sentiments via Linear SVM, Multinomial Naïve Bayes, Stochastic Gradient Descent, Random Forest, Logistic regression, and Decision Tree. The best classification results were obtained by SVM with an accuracy of 94.02% on musical domain. Singla et al. (2017) have performed sentiment analysis on Amazon mobile phone reviews and in their study, they have categorized text into positive and negative polarity, and have also included sentiments of anger, anticipation, fear, joy, sadness, disgust, surprise, and trust. The classification is done through SVM resulting in an accuracy of 84.85%. Moreover, Samsung brand received the most positive feedback from customers. These results are useful for manufacturers as they can work on the feedback to improve the quality of product.

In more recent times, demonetization was announced in India, where the Government of India demonetized ₹ 500 and ₹ 1000 banknotes. Many authors have used the enormous data that originated during this period for sentiment analysis. They predicted the pros and cons of demonetization in India and analyzed the support given to the Indian government by the people. Singh et al. (2017) proposed an approach to evaluate the effect of demonetization on people all over the world. They applied a lexicon-based approach Valence Aware Dictionary and Sentiment Reasoner (VADER) for sentiment analysis on tweets extracted from Twitter. They also performed a retweet analysis in which they predicted the tweet, which could be retweeted again by using various machine learning algorithms. Experimental results show that SVM obtained finest accuracy in the case of unigram, whereas unigram-bigram feature extraction and Classification and Regression Trees (CART) gave good results for bigram feature extraction. This wide range of applications motivates us to dig deep into this area for further research.

1.2 Review methodology

The following journal databases have been explored to conduct this survey:

- Springer Link
- Science Direct
- IEEE Xplore Digital Library
- Google Scholar
- ACM Digital Library
- Wiley Online Library

We have studied over 200 research papers from the above journals, IEEE, Springer, and ACM International Conferences, and Springer Book Chapters, and have shortlisted 130 research papers on sentiment analysis, which focuses on deep learning techniques only. The search

Table 2 Count of number of reviewed articles

S. no.	Articles in journals/conference proceedings/book chapters	Count
1	Neurocomputing	11
2	Expert Systems with Applications	9
3	ACM Transactions on Asian and Low-Resource Language Information Processing	1
4	Knowledge-Based Systems	6
5	Image and Vision Computing	3
6	Journal of Parallel and Distributed Computing	1
7	Neural Processing Letters	1
8	Information Fusion	5
9	Artificial Intelligence Review	5
10	IEEE Intelligent System	2
11	Information Processing and Management	2
12	Cognitive Computation	3
13	Neural Networks	1
14	IEEE/ACM Transactions on Audio, Speech, and Language Processing	1
15	IEEE Transactions on Knowledge and Data Engineering	2
16	Journal of Computational Science	1
17	IEEE Transactions on Affective Computing	1
18	IEEE Transactions on Multimedia	1
19	IEEE Transactions on Neural Networks and Learning Systems	1
20	ACM Transactions on Information Systems (TOIS)	1
21	Journal of King Saud University—Computer and Information Sciences	1
22	Decision Support Systems	2
23	Future Generation Computer Systems	2
24	Computer Vision, Springer	1
25	Language Resources and Evaluation, Springer	1
26	Journal of the American Society for Information Science and Technology	1
27	Theoretical Computer Science	1
28	Applied Intelligence, Springer	1
29	IEEE Access	4
30	Data mining and Knowledge Discovery, Wiley	2
31	Deep Learning in Natural Language Processing, Springer	1
32	IEEE Conference Proceedings	39
33	ACM Conference Proceedings	16
	Total	130

terms or keywords used in these databases include the following: Sentiment analysis, Opinion Mining, Sentiment analysis with deep learning, opinion mining with deep learning, Online Social Networks, Social Media, Deep Learning. The distribution of articles on sentiment analysis with deep learning is shown in Table 2.

An year-wise analysis for sentiment analysis using deep learning approaches is shown in Fig. 1. From Fig. 1, it can be seen that the research in this area is increasing immensely, and from the year 2015 onwards, the number of articles published in this area are growing.

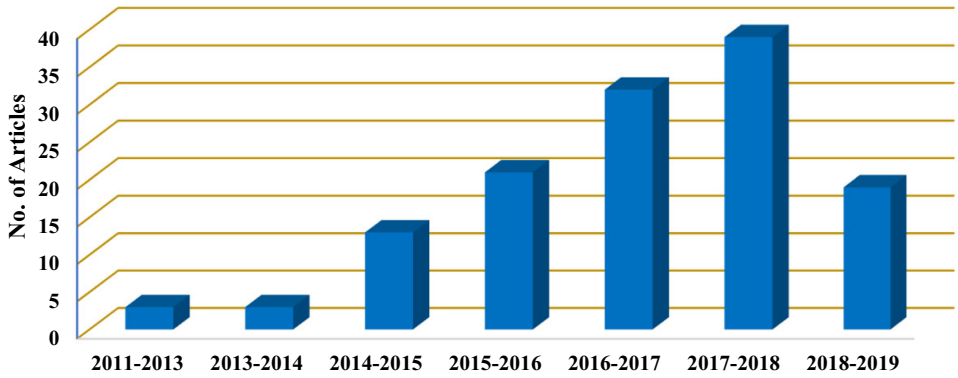


Fig. 1 Year-wise distribution of articles

Hence, our objective is to study the deep learning approaches focussed on sentiment analysis.

1.3 Earlier state-of-the-art surveys

This section discusses how our survey is different from the earlier state-of-the-art surveys. Most of the existing surveys have focused on specific areas in sentiment analysis like subjectivity detection (Chaturvedi et al. 2018) aspect extraction (Rana and Cheah 2016), social context (Sánchez-rada and Iglesias 2019), multimodal analysis (Poria et al. 2017a), information fusion (Balazs and Velásquez 2016), different languages and genre in sentiment analysis (Rani and Kumar 2019), multilingual sentiment analysis (Lo et al. 2017), and lexicon-based versus machine learning-based approaches for sentiment analysis (Hemmatian and Sohrabi 2017). Different from these surveys, this review aims to cover the significant and widespread approaches which are introduced recently in the field of sentiment analysis using deep learning. However, earlier methods are still included to give a complete view of the sentiment analysis research. The previous surveys fail to provide a detailed discussion and a comparative analysis of deep learning-based approaches for sentiment analysis, which this review aims to address. Hence, we outline the existing methods by presenting a taxonomy which explores the power of deep learning approaches and discusses how these approaches improve the performance of sentiment analysis.

This survey provides a comprehensive review of the existing literature on sentiment analysis with deep learning architecture. The significant contribution of this survey can be summarized as:

- Outlined a taxonomy for sentiment classification (Fig. 3), which includes methods using handcrafted and machine-learned feature-based approaches.
- Discuss the architecture of various deep learning-based approaches to summarize the notable work and highlight the popular deep learning-based architecture used by various researchers for sentiment analysis.
- Present the vital sentiment analysis tasks and identify why deep learning models are increasingly applied to them.
- Identify the popular sentiment analysis datasets, the accuracy obtained on them, and discuss the need for creating own corpus for sentiment analysis.

The organization of the survey is shown in Fig. 2. Section 2 discusses the taxonomy and tasks related to sentiment analysis. Section 3 gives a detailed description of the deep learning models along with an application-wise comparison of deep learning approaches, drawbacks

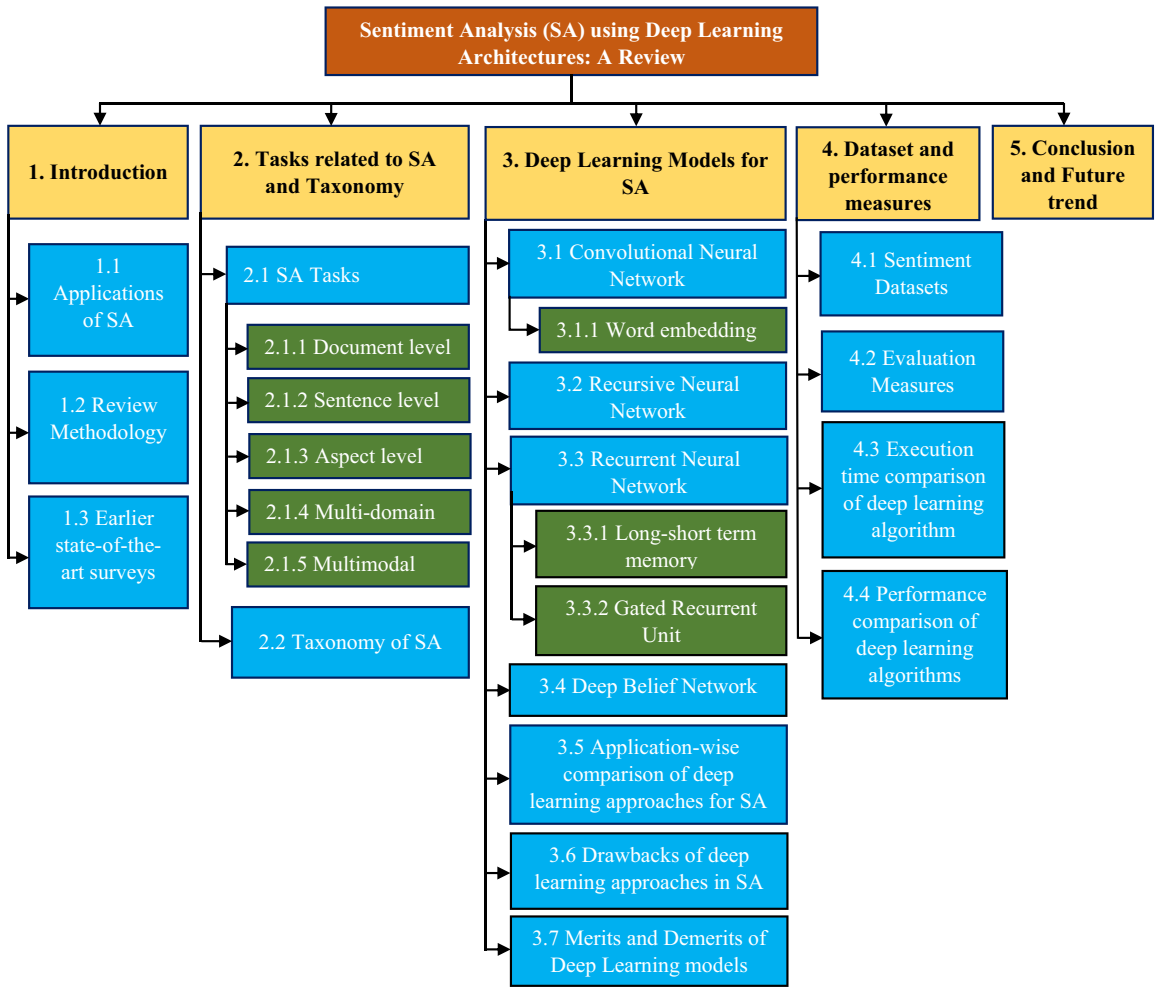


Fig. 2 Organization of the survey

of these approaches, and merits and demerits of popular deep learning models. Section 4 discusses sentiment analysis datasets and performance measures. Finally, Sect. 5 presents the conclusion and future trend.

2 Tasks related to sentiment analysis and taxonomy

This section discusses about the popular sentiment analysis tasks. Further, we have made considerable efforts to identify a taxonomy for sentiment analysis.

2.1 Sentiment analysis tasks

The sentiment analysis tasks can be categorized into two parts (Ravi and Ravi 2015): core (or major) tasks which are referred as the basic sentiment analysis tasks, and second category includes the sub-tasks that are called sub-categories of the major tasks. The core sentiment analysis tasks include document-level sentiment classification, sentence-level sentiment classification, and aspect-level sentiment classification, and sub-tasks include multi-domain sentiment classification and multimodal sentiment classification (Soleymani et al. 2017). Apart from them, various other sub-tasks of sentiment analysis, which are getting a lot of

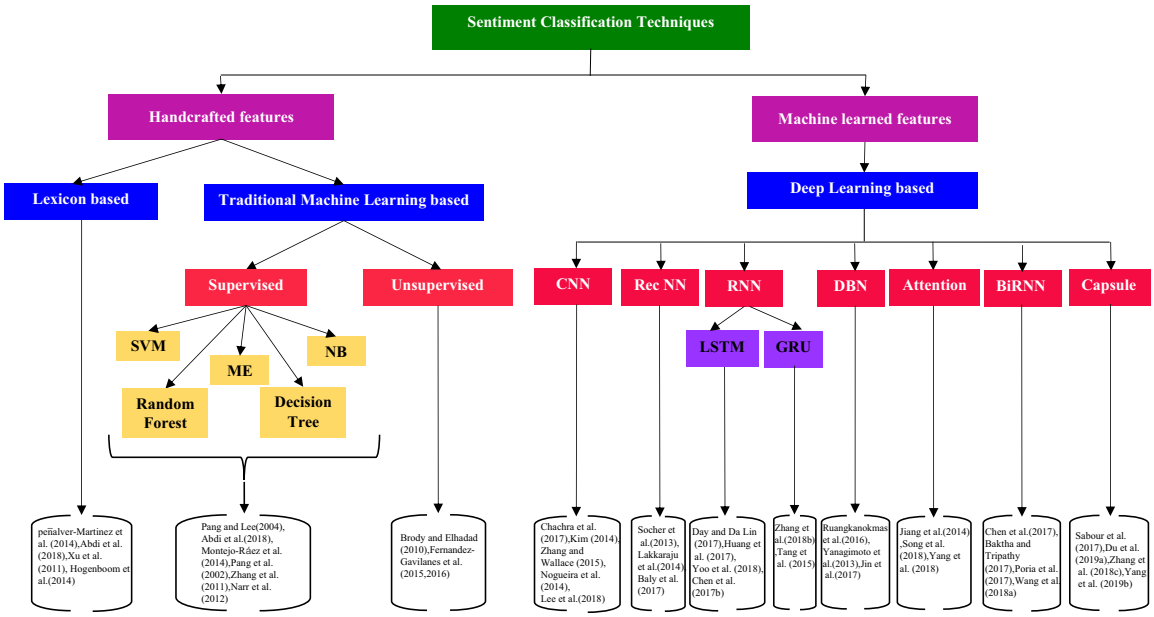


Fig. 3 Taxonomy of sentiment analysis

attention from the researchers are also identified. The various tasks and sub-tasks of sentiment analysis are discussed in the subsequent sections.

2.1.1 Document-level sentiment classification

This level of classification takes the whole document as a primary unit of information focusing on one topic or object. The document is further categorized into positive polarity or negative polarity. Thus, an overall sentiment of text can be generated. Yang et al. (2016) proposed a hierarchical attention network model that focuses on vital content for constructing the document representation. Experimental results on six popular text-based reviews demonstrate that the proposed model outperformed the state-of-the-art results by a significant margin as it can capture the insights about the structure of the document. A major challenge in document-level sentiment classification is to model long texts for generating semantic relations between sentences. This problem was handled by Huang et al. (2018), who proposed a model called SR-LSTM in which the first layer used LSTM to learn the sentence vectors, and second layer encodes the relations between the sentences. A hybrid approach of RBM and Probabilistic Neural Network (PNN) is proposed by Ghosh et al. (2017) in which RBM is used for dimensionality reduction, and PNN performs sentiment classification. The experiment is conducted in four steps: Initially, multi-domain data is collected containing reviews on Movies, Books, DVDs, Electronics, and Kitchen appliances. Next, the data is pre-processed using tokenization, stemming, and stop-word removal. The third step includes dimensionality reduction on the dataset using RBM, and finally, PNN was use for binary sentiment classification. The proposed approach gave better results on the five datasets compared to the state-of-the-art methods.

2.1.2 Sentence-level sentiment classification

The disadvantage of document-level sentiment classification is that it is difficult to extract the different polarity or sentiment about distinct entities separately. Hence, in sentence-level

sentiment classification, a sentence is classified into subjective type or objective type. A *subjective* statement expresses an opinion towards an entity. For example, "*I got a beautiful bag*", signifies positive polarity about *bag*. Hence, it is considered as a subjective statement that can be further classified into different polarities. On the other hand, factual statements are termed as *objective* statements. A statement like "*The bottle is blue in color*", displays no sentiment, so it is categorized as an objective statement.

Zhao et al. (2017) proposed a framework called Weakly-supervised Deep Embedding (WDE), which employs review ratings to train a sentiment classifier. They used CNN for constructing WDE-CNN and LSTM for constructing WDE-LSTM to extract feature vectors from review sentences. The model was evaluated on Amazon dataset from three domains: digital cameras, cell phones, and laptops. The accuracy obtained on WDE-CNN model was 87.7%, and on WDE-LSTM model was 87.9%, which shows that deep learning models gives highest accuracy as compared to baseline models. Xiong et al. (2018b) developed a model called Multi-level Sentiment-enriched Word Embedding (MSWE), which uses a Multi-layer perceptron (MLP) to model word-level sentiment information and CNN to model tweet-level sentiment information. The model also learns sentiment-specific word embeddings, and SVM is used for sentiment classification. It was evaluated on SemEval2013 dataset and Context-Sensitive Twitter (CST) dataset, which are the benchmark datasets for sentiment classification task. The F1 score obtained in the SemEval2013 dataset was 85.75, and on CST dataset was 81.34.

2.1.3 Aspect-level sentiment classification

Aspect level sentiment analysis is commonly called feature-based sentiment analysis or entity-based sentiment analysis. This sentiment analysis task includes the identification of features or aspects in a sentence (which is a user-generated review of an entity) and categorizing the features as positive or negative. The sentiment-target pairs are first identified, then they are classified into different polarities, and finally, sentiment values for every aspect are clubbed. Peng et al. (2018) studied the Chinese aspect targets at three granularity levels: radical, character, and word by proposing a model called Aspect Target Sequence Model for Single Granularity (ATSM-S). The previous work was related to processing only one aspect at a time, so they addressed this issue and presented an approach to process two aspects at a time by focusing on the aspect target itself.

Recently, attention-based LSTM mechanisms are being used for aspect-based sentiment analysis. Wang et al. (2016a) proposed an attention-based LSTM model, which can focus on different parts of a sentence when various aspects are concerned. The attention weights are computed by concatenating aspect vector into the sentence hidden representation (AE-LSTM model) or by appending aspect vector embedding into each word input vector (ATAE-LSTM model). Experimental results demonstrate that both the proposed models achieved superior performance over the baseline models, which shows that attention-based LSTM models boost the performance of aspect-based sentiment analysis models. Yu et al. (2019) proposed a framework using Bi-LSTM and multi-layer attention networks for aspect and opinion terms extraction. Al-Smadi et al. (2018) proposed a bi-LSTM with CRF model for aspect opinion target expressions (OTEs) extraction, along with aspect-based LSTM where the aspect OTEs are treated as attention expression for aspect sentiment polarity classification. Ma et al. (2018) proposed a two-step attention architecture, which attends words of the target expression along with the whole sentence. The author also applied extended LSTM, which can utilize external knowledge for developing a common-sense system for target aspect-based sentiment analysis.

The initial systems were not able to model different aspects in a sentence and do not explore the explicit position context of words. Hence, Ma et al. (2019) developed a two-stage approach that can handle the above problems. In Stage-1, position attention model is introduced for modelling the aspects and its neighboring context words. In Stage-2 multiple aspect terms within a sentence are modelled simultaneously. The most recent approach is proposed by Yang et al. (2019a), which replaces the conventional attention models with coattention mechanism by introducing a Coattention-LSTM network that can model the context-level and target-level attention alternatively by learning the non-linear representations of the target and context simultaneously. Thus, the proposed model can extract more effective sentiment features for aspect-based sentiment analysis.

2.1.4 Multi-domain sentiment classification

The word *domain* is referred as a set of documents that are related to a specific topic. Multi-domain sentiment classification focuses on transferring information from one domain to the next domain. The models are first trained in source domain; the knowledge is then transferred and explored in another domain. Dragoni and Petrucci (2017) incorporated word embeddings with a deep learning model for implementing a NeuroSent tool to build a multi-domain sentiment model. Yuan et al. (2018) proposed a Domain Attention Model (DAM) for modeling the feature-level tasks using attention mechanism for multi-domain sentiment classification. DAM is composed of two modules: domain module and sentiment module. The domain module predicts the domain in which text belongs using bi-LSTM, and sentiment module selects the important features related to the domain using another bi-LSTM with attention mechanism. The vector thus obtained from the sentiment module is fed into a softmax classifier to predict the polarity of the texts. The author used Amazon multi-domain dataset containing reviews from four domains, and Sanders Twitter Sentiment dataset containing tweets about four different IT companies. The proposed model was compared with traditional machine learning approaches, and results show that the model performed well for multi-domain sentiment classification.

2.1.5 Multimodal sentiment classification

Different people express their sentiments or opinions in different ways. Earlier, the text was considered as the primary medium to express an opinion. This is known as a unimodal approach. With the advancement of technology and science, people are now shifting towards visual (videos, images, or clips) and audio (speech) modalities to express their sentiments. Combining or fusing more than one modalities for detecting the opinion is known as multimodal sentiment analysis. Hence, researchers are now focusing on this direction for improving the sentiment classification process.

Chen et al. (2018) proposed a Weakly-Supervised Multi-modal Deep Learning (WS-MDL) model to predict multimodal sentiments for tweets. The model uses CNN and Dynamic CNN (DCNN) to calculate multimodal prediction scores and sentiment consistency scores. Due to the enormous data available on social media in different forms like videos, audios, photos for expressing sentiment on social media platforms, the conventional approach for text-based sentiment analysis was progressed into compound models of multimodal sentiment analysis. Hence, mining the opinions expressed in different modalities became a crucial approach. Poria et al. (2016a) proposed a novel methodology for merging the affective information extracted from audio, visual, and textual modalities. They discussed how different modalities

were combined together to improve the overall sentiment analysis process. Poria et al. (2018) explored three deep learning architectures for unimodal, bimodal, and multimodal (trimodal) sentiment classification. The experimental results showed that bimodal and trimodal models have shown better accuracy as compared to unimodal models, which shows the importance of using features from all the modality for enhancing the performance of sentiment analysis models. For more information on multimodal sentiment analysis, the following popular works can be referred (Soleymani et al. 2017; Chen et al. 2018; Poria et al. 2016b; Shah et al. 2016; Zhang et al. 2018a; Agarwal et al. 2019).

Table 3 discussed the deep learning models applied to various sentiment analysis (SA) tasks. From Table 3, we can conclude that deep learning models are gaining immense popularity for various SA tasks. The popular models applied to SA tasks can be summarised as:

- For document-level sentiment classification, CNN followed by LSTM has shown the highest accuracy on the various dataset.
- For sentence-level sentiment classification and aspect-level sentiment classification, researchers have majorly focused on RNN (particularly LSTM).
- For multi-domain sentiment classification, LSTM has given good results, and for multimodal sentiment classification, CNN and RNN are popular deep learning models. Hence, RNN models are the most sought-after and popular choice for sentiment analysis among researchers.
- Further, we can see that LSTM is popularly applied for text-based sentiments, and CNN models have shown good results for image sentiment. For multimodal data, CNN + LSTM followed by fusion becomes the desired approach.

However, the choice of a specific deep learning model may still depend on the various number of factors like the amount of data available, the number of hidden units (nodes) required for the problem, etc.

We found that the popular languages used for sentiment analysis are: English, Chinese, and Arabic. The detailed statistics about different types of languages used by various researchers are shown in Table 4 below.

Apart from the above-mentioned sentiment analysis tasks, we have identified various other sub-tasks that are gaining a lot of attention. This includes sarcasm detection (Halin 2017), sentiment summarization (Abdi et al. 2018), irony detection (van Hee et al. 2018), implicit sentiment detection (Chandankhede et al. 2016), temporal tagging (Hafez et al. 2017), stance detection (Krejzl et al. 2017), and emotion analysis (Hakak et al. 2017).

2.2 Taxonomy of sentiment analysis

Research in the field of sentiment analysis is taking place for several years. Initially, handcrafted features were used for various classification tasks. Some examples of handcrafted features are shown in Table 5. Lexicon based methods use handcrafted features and depend on sentiment lexicons, which are the collection of lexical units and their sentiment orientation.

On the other hand, machine-learned features can be categorized into traditional machine learning-based approaches and deep learning-based approaches. Machine learning-based methods include Support Vector Machine (SVM), Naïve Bayes (NB), Maximum Entropy (ME), Decision tree learning, and Random Forests. They are further categorized into supervised and unsupervised learning methods. A taxonomy of various approaches for sentiment analysis can be developed, as shown in Fig. 3. From the figure, it becomes evident that the number of established approaches for supervised learning is more as compared to unsu-

Table 3 Deep learning model in SA

Refs.	Sentiment analysis task	Language	Sentiment	Deep learning models	Dataset	Accuracy or F1 score (%)
Uysal and Murphey (2017)	Document-level sentiment classification	English	Text	CNN, LSTM, CNN + LSTM	IMDB (Maas et al. 2011) Sentiment140 (Go et al. 2009) Nine Public Sentiment reviews (Whitehead and Yaeger 2009) Amazon Multi-domain dataset (Blitzer et al. 2007)	IMDB: 89.1 Sentiment140: 71.5 Nine Public: 77.1 Multi-domain dataset: 85.4
Huang et al. (2018)	Document-level sentiment classification	English	Text	LSTM	Yelp 2014 (https://www.yelp.com/dataset/challenge) Yelp 2015 (https://www.yelp.com/dataset/challenge) IMDB (Maas et al. 2011)	Yelp 2014: 63.9 Yelp 2015: 63.8 IMDB: 44.3

Table 3 continued

Refs.	Sentiment analysis task	Language	Sentiment	Deep learning models	Dataset	Accuracy or F1 score (%)
Ghosh et al. (2017)	Document-level sentiment classification	English	Text	RBM and PNN	Movie review (MOV) (Pang and Lee 2004) Multi-domain dataset: Books (BOO), DVDs, Electronics (ELE), and Kitchen appliances (KIT) (Blitzer et al. 2007)	MOV: 80.8 BOO: 81 DVD: 81.1 ELE: 80.1 KIT: 80.2
Shi et al. (2017)	Document-level sentiment classification	Chinese	Text	Hierarchical LSTM model	Crawl weibos from Sina Weibo (http://weibo.com)	90.8
Yanagimoto et al. (2013)	Sentence classification	Japanese	Text	RBM	T&C news	–
Hassan and Mahmood (2018)	Sentence- level classification	English	Text	Joint CNN and RNN framework	IMDB (Maas et al. 2011) SSTb (Socher et al. 2013)	IMDB: 93.2 SSTb: 48.8 (fine-grained) and 89.2 (binary)
Hassan and Mahmood (2017a)	Sentence- level classification	English	Text	CNN and BRNN	IMDB (Maas et al. 2011) SSTb (Socher et al. 2013)	IMDB: 93.4 Stanford: 48.9 (Fine-grained) 89.6 (Binary)
Zhao et al. (2017)	Sentence-level classification	English	Text	CNN + LSTM	Review from Amazon on digital cameras, cell phones and laptops	WDE-CNN: 87.7 WDE-LSTM: 87.9

Table 3 continued

Refs.	Sentiment analysis task	Language	Sentiment	Deep learning models	Dataset	Accuracy or F1 score (%)
Day and Da Lin (2017)	Sentence- level classification	Chinese	Text	Bidirectional-LSTM	Google Play consumer review	94
Xiong et al. (2018b)	Sentence- level classification	English	Text	MLP + CNN	SemEval2013 (Nakov et al. 2013) Context-Sensitive Twitter (CST) (Ren et al. 2016)	(F1 Score) 85.75 on SemEval 2013 And 81.34 on CST
Jin et al. (2017)	Sentence- level classification	English	Text	DBN with Delta Rule based on RBM	Stanford Twitter Sentiment dataset (STS-T) (Saif et al. 2013) STS-G (Saif et al. 2013) Health Care Reform (HCR) (Saif et al. 2013) Sentiment Strength Twitter (SST) (Saif et al. 2013) Full Twitter data (FT) (Kharde and Sonawane 2016) Game Tweets (GT)	Highest Average accuracy 67.44
Li et al. (2014)	Sentence- level classification	Chinese	Text	Recursive Neural Deep Model (RNDM)	Chinese Sentiment Treebank	90.8

Table 3 continued

Refs.	Sentiment analysis task	Language	Sentiment	Deep learning models	Dataset	Accuracy or F1 score (%)
Vateekul and Koomsubha (2016)	Sentence- level classification	Thai	Text	LSTM and Dynamic Convolutional Neural Network (DCNN)	Twitter	LSTM: 75.30 DCNN: 75.35
Socher et al. (2013)	Sentence- level classification	English	Text	Recursive Neural Tensor Network (RNTN) and MV-RNN	Sentiment Treebank	RNTN: 2 way: 80.7 5 way: 87.6 MV-RNN: 2 way: 86.8 5 way: 78.7
Wu and Chi (2017)	Sentence- level classification	English	Text	qLSTM-RecNN	Stanford Sentiment Treebank (SST) (Socher et al. 2013) and SICK data set (Marelli et al. 2014)	**SST**: 86.6 (Binary) 49.4 (Fine grained) **SICK**: 87.28
Al-Smadi et al. (2017)	Aspect level sentiment classification	Arabic	Text	RNN with word embedding and SVM	Arabic Hotels' reviews dataset (Al-Smadi et al. 2016; Pontiki 2016)	RNN: 87
Peng et al. (2018)	Aspect level sentiment classification	Chinese and English	Text	LSTM + Attention mechanism	Chinese dataset (CD) SemEval2014 (SE) (Pontiki et al. 2014)	CD: 85.95 SE: 75.39
Tay et al. (2017)	Aspect level sentiment classification	English	Text	Dyadic Memory Networks	Customer reviews for laptop, restaurants, SemEval2014, Tweets from SemEval 2016, Online Debates	(F1 score) 69.2

Table 3 continued

Refs.	Sentiment analysis task	Language	Sentiment	Deep learning models	Dataset	Accuracy or F1 score (%)
Lakkaraju et al. (2014)	Aspect level sentiment classification	English	Text	Joint Multi-Aspect Sentiment Model (JMAS) + RNTN	Beer reviews and camera reviews	Beer: 77.04 Camera: 81.02
Yuan et al. (2018)	Multi-domain sentiment classification	English	Text	bi-direction LSTM with attention mechanism	Amazon multi-domain dataset (Amazon) (Blitzer et al. 2007) Sanders Twitter Sentiment Dataset (Sanders)	(Average accuracy) Amazon: 87.69 Sanders: 86.32
Nozza et al. (2016)	Multi-domain sentiment classification	English	Text	Autoencoder [marginalized Stacked Denoising Autoencoder (mSDA)] and decoder	Reviews from Amazon multi-domain dataset (Blitzer et al. 2007)	87
Dragoni and Petrucci (2017)	Multi-domain sentiment classification	English	Text	RNN	Dranziera dataset (Dragoni et al. 2016)	(Average F1 score) 84.460
Poria et al. (2016b)	Multimodal sentiment classification and emotion detection	English and Spanish	Videos	Temporal CNN, RNN	MOUD dataset Youtube (Morency et al. 2011) ICT-MMMO	96.55
Chen et al. (2018)	Multimodal sentiment classification	Chinese	Image + Text + Emoticon	Visual sentiment prediction-AlexNet based CNN Textual sentiment prediction-Word2Vec and DCNN	Microblogs crawled from Sina Weibo (http://weibo.com)	69.5 (approx.)

Table 3 continued

Refs.	Sentiment analysis task	Language	Sentiment	Deep learning models	Dataset	Accuracy or F1 score (%)
Chen et al. (2017a)	Multimodal sentiment classification	English	Image + Text	Deep fusion CNN	VSO (Borth et al. 2013): Flickr Images MVSO-EN (Jou et al. 2015): the dataset includes concepts related to emotions expressed in images	VSO: 84.7 MVSO-EN: 73.7

Table 4 Count of articles according to the languages in sentiment analysis by various researchers

Language	English	Chinese	Arabic	Japanese	Vietnamese	Thai	Tibetan	Persian	Spanish	Punjabi	Chinese and English	Korean and English
Count of articles	57	18	4	2	2	1	1	1	1	1	5	2

Table 5 List of handcrafted features

Handcrafted feature	Description
PoS (Part-of-speech) (Pang and Lee 2004; Abdi et al. 2018; Poria et al. 2016c; Peñalver-Martinez et al. 2014; Montejo-Ráez et al. 2014)	PoS reads the input text and label each word with parts of speech such as nouns, adverbs, verbs, adjectives, etc.
Word n-gram, character n-gram (Pang and Lee 2004; Peñalver-Martinez et al. 2014; Pang et al. 2002; Zhang et al. 2011; Narr et al. 2012)	Word n-grams is a combination of a contiguous sequence of *n* words in the input text sequence. They are usually categorized into unigram (size 1), bigram (size 2), trigram (size 3). Similarly, character n-grams are the combination of a continuous sequence of *n* characters in the input text sequence
Hashtags (Xiong et al. 2018; Montejo-Ráez et al. 2014; Stojanovski et al. 2018)	They are majorly used for Twitter sentiment analysis by counting the positive and negative hashtag tokens in the tweet
Emoticons (Xiong et al. 2018; Montejo-Ráez et al. 2014; Narr et al. 2012; Stojanovski et al. 2018)	They are majorly used for Twitter sentiment analysis by counting the positive and negative emoticons in the tweet
Punctuation (Abdi et al. 2018; Abdi et al. 2019; Xu et al. 2011; Abbasi et al. 2008)	Occurrence of punctuation marks e.g., question mark (?), exclamation mark (!), colon (::) are considered as features
Negations (Abdi et al. 2018; Abdi et al. 2019; Zhang et al. 2017; Liu et al. 2017)	Negation words like don't, didn't, not, etc. can change the polarity of the sentence from positive to negative and vice versa
Sentiment lexicon (Abdi et al. 2018; Peñalver-Martinez et al. 2014; Abdi et al. 2019; Zhang et al. 2017)	Sentiment lexicon contains a list of words which expresses a positive or negative sentiment. The final sentiment is calculated by averaging the #positive tokens and #negative tokens. e.g., MPQA lexicon (Kiritchenko et al. 2014), SentiwordNet 3.0 (Baccianella et al. 2010), Sentiment140 lexicon (Kiritchenko et al. 2014), NRC Hashtag Sentiment Lexicon (Mohammad et al. 2013) etc.

pervised learning-based approaches. Figure 3 also gives an idea about a large number of articles published using deep learning-based approaches. These approaches include Convolution Neural Network, Recursive Neural Network, Recurrent Neural Network (which includes LSTM and GRU), Deep Belief Networks, Attention-based networks, Bi-directional Recurrent Neural Network, and capsule network.

Cambria (2016) presents a categorization for the tasks related to affective computing and sentiment analysis. The primary tasks include emotion recognition and polarity detection. The general categorization of these tasks based on existing approaches is: knowledge-based techniques, statistical techniques, and hybrid approaches. The knowledge-based techniques include several lexicons from which certain affective words are extracted for text classification. This approach is not able to handle different nuances and results in poor representation of emotions or sentiments due to the presence of certain linguistic rules. This motivated researchers to develop statistical techniques which are based on machine learning and deep learning approaches that can learn the complex features from the data, thus improving the sentiment analysis process. However, these approaches require lots of data to sufficiently train themselves. Hence, hybrid strategies that combine knowledge-based techniques and statistical techniques are popularly used for performing emotion recognition and polarity detection.

The popularity of sentiment analysis has attracted many researchers to work in this area. A lot of existing research in this field focuses on machine learning based methods. Deep learning is one of the fastest-growing areas which falls under the category of machine learning. Hence, deep learning is considered as a subject for study as it has potential benefits over other methods due to the following reasons:

- Traditional approaches, like lexicon-based approaches, use handcrafted features, which is a time-consuming and tedious process. Moreover, they are not able to generalize well for other domains or areas. Even in traditional machine learning approaches, feature engineering and feature extraction are the most time-consuming process. Hence, deep learning reduces the burden of feature design as when the network learns, it automatically creates the required features for the classification process.
- At present, a massive amount of data is being generated. According to Twitter, an average of around 6000 tweets are produced per second, which means about 200 billion tweets are being generated per year. Hence, with such a massive quantity of data, traditional machine learning-based approaches fail to perform. On the contrary, deep learning models outperform the machine learning approaches as they can be trained to learn more features with large datasets. This justifies the fact that deep learning models show improved performance than traditional machine learning models.
- The multiple layers of deep learning architecture can capture non-linear and intricate patterns in the data.
- Deep learning architectures can be adapted to other domains like Image Processing, Medical Image Segmentation, Internet of Things (IoT), Speech Recognition.

3 Deep learning models for sentiment analysis

Recent years have shown a trend of deep learning models applied in the field of natural language processing (NLP). Deep neural networks (DNNs) are made up of artificial neural networks having multiple hidden layers between the input layer and the output layer. This section provides a detailed discussion about some of the most popular deep learning mod-

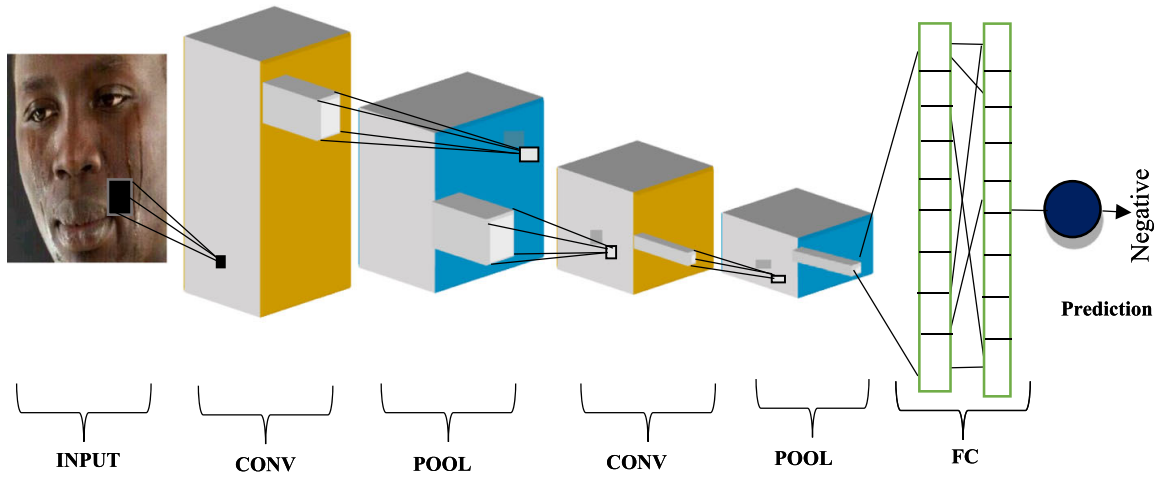


Fig. 4 CNN architecture

els like CNNs, Rec NNs, RNNs, and deep belief networks along with a brief overview of Attention-based networks, Bi-directional RNNs, and capsule networks. Further, we provide an application-wise comparison of deep learning approaches, along with their merits and demerits.

3.1 Convolutional neural networks (CNNs)

CNNs belong to the class of neural networks and have shown significant success and innovation in computer vision and image processing. The fundamental architecture of CNNis displayed in Fig. 4. As evident from the figure, CNN consists of various layers, such as the input layer, convolutional layer, pooling layer, and fully connected layer. The task of the input layer is to take the pixel value of the image as an input. Next, convolution layer (CONV) has the responsibility to produce output based on its kernel or filter values. The output obtained through a convolution operation, and Pooling Layer (POOL) is used to reduce the size of representation (dimensionality) and to speed up computation.

The most popular type of pooling is max pooling, in which maximum value from each window is taken. The Fully connected layer (FC) connects every neuron in this layer to all the activations of previous layer, as seen in ordinary neural networks. More and more researchers are actively using CNNs in the field of sentiment analysis. The most popular CNN model for sentence-level sentiment classification is the work done by Kim (2014). The author conducted an experiment with CNN built on top of pre-trained word2vec. The experimental results show that pre-trained vectors can serve as an excellent feature extractor for tasks related to NLP using deep learning. Motivated by these results, Zhang and Wallace (2015) discussed an architecture for sentence classification using one-layer CNN. They explored how the performance of a model can be affected by changing its configuration (hyperparameters, filter size, regularization parameters, etc.). Figure 5 illustrates the architecture proposed by Zhang and Wallace (2015). The tokenized sentence of length s is given as an input to the network, and it is converted into a sentence matrix by following the work of Collobert and Weston (2008) which applies a look-up table concept to generate the sentence matrix. The dimensionality of the matrix is $s * d$, where d represents dimensionality of word vectors. Hence, sentence matrix can now be treated as an input image on which convolution is performed using linear filters to generate the feature maps. The height of the filter is referred as region size of the filter. For pooling operation, 1-max pooling is performed on each feature map.

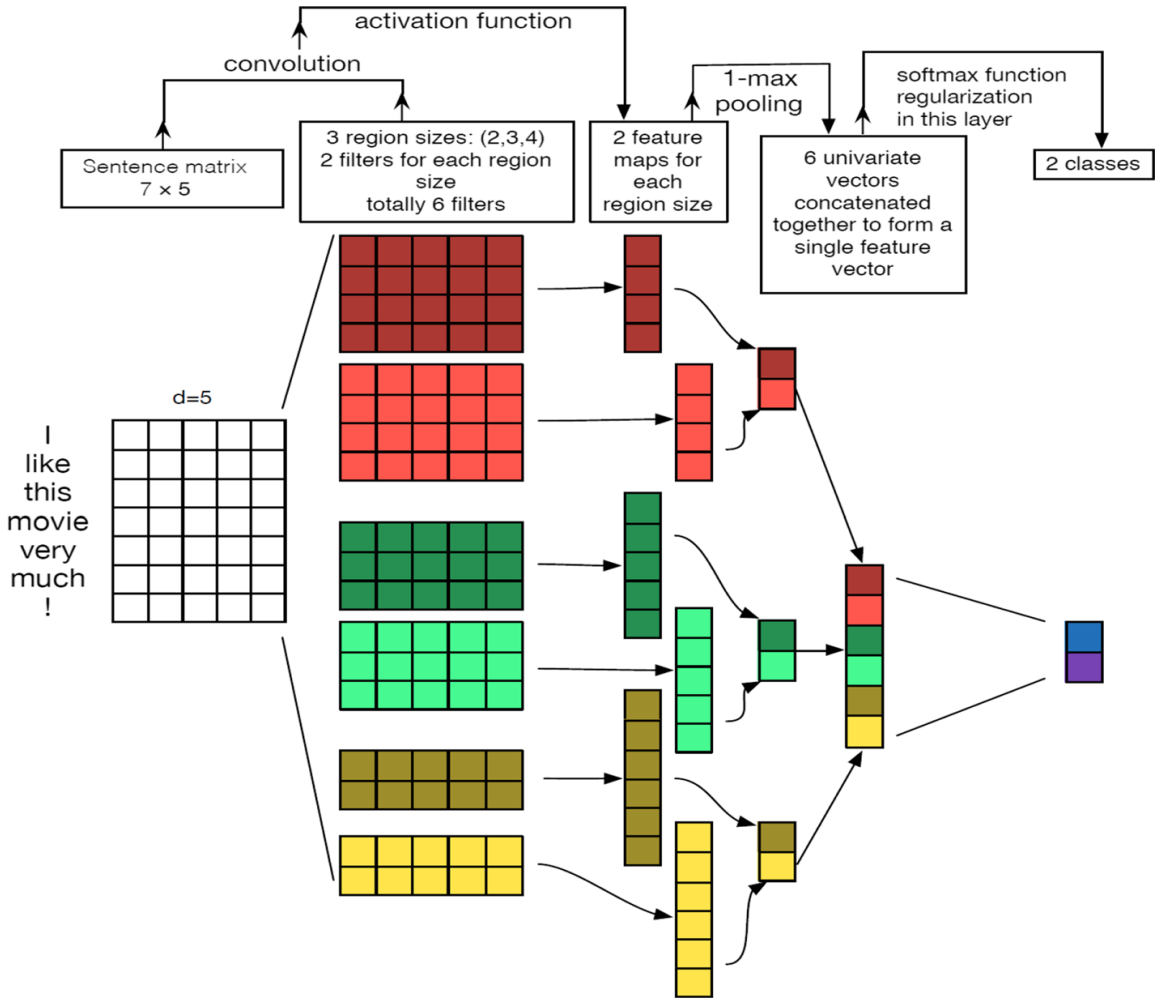


Fig. 5 CNN architecture for sentiment classification (Zhang and Wallace 2015)

3.1.1 Word embedding

Inspired by the ideas, architecture, and results of CNNs in the field of computer vision, CNNs are gaining popularity in the domain of NLP too. In NLP related tasks, an input layer consists of the matrix representation of sentences or documents, instead of the image pixels. Each row of the matrix is a vector representation of either a word or a character. These vectors are called word embedding or character embedding. The earlier approach used to represent the vocabulary of a document was one-hot encoding. The problem with this approach is that the vector size increases with the corpus size. Moreover, this encoding is not able to capture the relationship between words. Hence, word embeddings were developed as one of the most popular techniques for representing the vocabulary of a document. They contain a set of feature selection methods or a set of language models that maps the textual word into its equivalent dense and low-dimensionality vector representations. They can capture context of the word and can provide information about relation of a word with other words. Hence, meaning of a word can be predicted accurately as it can capture syntactic and semantic information about the words.

Word2Vec (Mikolov et al. 2013) is one of the famous technique for learning word embeddings as they use shallow neural network for processing a text before passing it into a deep learning algorithm. The embeddings can be obtained using Skip Gram model and Common Bag of words (CBOW) model. The CBOW model predicts the current word from surrounding

context words, whereas Skip Gram model predicts the surrounding context words from the current word. The words are mapped into a word matrix and are converted into vectors in an n-dimensional vector space by representing similar words near to each other. Similarly, Global Vectors (GloVe) (Pennington et al. 2014) generates the vector encoding of a word. The advantage of Glove model is that it can be trained quickly on more data as the implementation can be parallelized. On the other hand, instead of learning the embedding of the full word, char2vec (Cao and Rei 2016) can learn embedding associated with each character of a word.

More recently, many researchers are coming up with new approaches on word representation for sentiment analysis. The traditional word embedding methods learn word distributions that are independent of any specific task. For sentiment analysis, this can be overcome by utilizing the prior knowledge that is available in the form of sentiment labels, or opinionated words from sentiment lexicons. Li et al. (2017a) proposed a framework which combines different levels of prior knowledge into the word embeddings for sentiment analysis. Experimental results on real-world data demonstrate that the proposed word representation method improves the performance of sentiment analysis systems when compared to baseline word embedding approaches. Hao et al. (2019) proposed a novel approach by applying stochastic embeddings for cross-domain sentiment classification, which preserves similarity structures in embedding space. Yu et al. (2018) proposed a model that learns sentiment embeddings by using sentiment intensity scores from sentiment lexicons. This improves the word vectors as they are semantically and sentimentally closer to similar words.

3.2 Recursive neural networks (RecNNs)

RecNN belongs to the category of the network, that learns a directed acyclic graph structure (e.g., tree structure). The weights are shared by using them recursively over an input, which is processed in hierarchical order. The network takes a structural representation of a sentence in the form of a parse tree with word vector representations at leaves, and recursively generates parent representations in a bottom-up manner. In this way, the tokens are combined to produce representations for phrases, and finally, a complete sentence is formed. The sentence representation obtained can then be used for final sentiment classification. In this way, RecNN can learn a hierarchical structure. Since each node of the tree has a distributed feature representation associated with it, we can add a softmax layer with each parent node for computing the label probabilities. These networks are generally abbreviated as RNNs, which also stands for Recurrent Neural Network. Hence, we have abbreviated Recursive Neural Networks as Rec NN to differentiate it from Recurrent Neural Networks (RNNs).

Figure 6 shows a Rec NN in which whenever a sequence of n-gram (e.g., *This Bag has Beautiful color*), is fed into the model, each word of n-gram sequence is represented as a d-dimensional vector. Let c_i and c_j be an n-dimensional vector representation of two-child nodes, as shown in Fig. 7, having an n-dimensional parent $p_{i,j}$. The parent vector must have the same dimensions so that they can be used as an input for the next composition. The parent vector $p_{i,j}$ is fed into a softmax classifier for computing the final label probabilities.

The advantage of Rec NNs is that they are powerful enough to learn a hierarchical network. Moreover, the vector representation of the words can be immediately used as feature inputs into a softmax classifier. The disadvantage of Rec NNs is that during the training phase itself, a tree representation of the input sequence needs to be known along with the *tag* of each word (In Fig. 7, the Parts of Speech tag is labeled with every word). Moreover, the structure of every input sequence changes for each training sample, thus making them hard for training purposes.

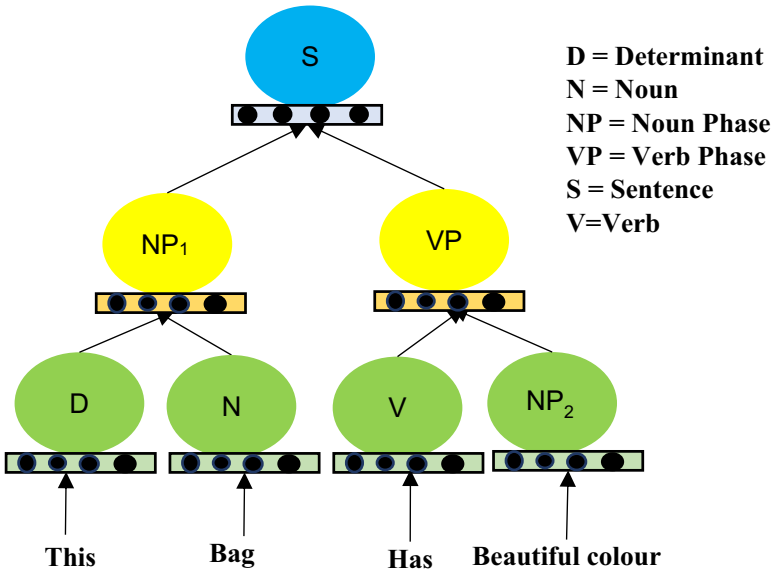


Fig. 6 Recursive neural network

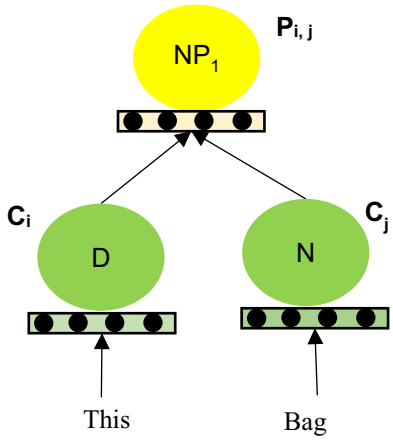


Fig. 7 One Rec NN which is reproduced for each pair of input vectors

Socher et al. (2013) discussed an approach for RecNNs, as shown in Fig. 8a. A tri-gram goes as an input to the network, and it is parsed into a tree-like structure where the leaves denote a word vector. The compositional function is denoted by $g(a, p_1)$, which computes the parent nodes (vectors) by moving in a bottom-up manner. These vectors are fed into the classifiers as features. They also introduce a new corpus called Stanford Sentiment Treebank (SSTb) which will help in analysing the compositional effects of sentiments expressed in any language. They applied Recursive Neural Tensor Network (RNTN) to capture these compositional effects and increase interactions between the input vectors, as shown in Fig. 8b.

RNN, Matrix–Vector RNN (MV-RNN), and Recursive Neural Tensor Network (RNTN) are the members of the family of Recursive Neural Models. Hence, all of them follow the bottom-up approach for computing a parent vector by applying a compositionality function and considering the vector nodes as features for a classifier. The major difference between a standard RNN and RNTN model is that the latter uses a tensor-based composition function for computing vectors of higher nodes, which allows the model to have more interactions between the nodes. Hence, RNTN is more powerful as it can combine the meaning of smaller constituents of a sentence more accurately than a standard RNN. Lakkaraju et al. (2014) focused on aspect-based sentiment analysis in which various aspects of a product or service

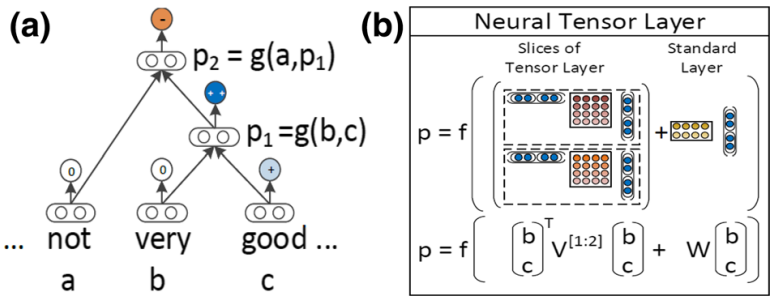


Fig. 8 a Recursive neural network model and **b** recursive neural tensor model with single layer (Socher et al. 2013)

are identified, and sentiments related to them are extracted using Joint Multi-Aspect Sentiment Model (JMAS). The results concluded that highest accuracy is shown by RNTN + JMAS followed by MV-RNN approach, which proves that simple RecNN is unable to capture the interactions between the constituents of sentences.

3.3 Recurrent neural networks (RNNs)

RNN is a variant of Rec NN and is used for modeling the sequential data. Sequential data is applied in a variety of applications. For example, in language translation, a sequence of sentence is translated from one language to another, in speech recognition an audio clip (sequence which plays over a time) is mapped into text script (sequence of words), and in video activity recognition, sequence of video frames are converted into text which describes the activity shown in video. The primary difference between RNN and Rec NN is that unlike Rec NN, RNN considers the time factor for processing the elements in a sequence. Thus, output in RNNs depends not only on the present input but also on the output computed from the previously hidden state of a network. RNNs stores the internal states of the inputs by processing each word in a sentence recurrently. Hence, to predict the next word in a sentence, RNN will store all the previous words and the relations between them.

Figure 9a shows an RNN with a loop that preserves all the information, whereas, in Fig. 9b, the unrolled version indicates that multiple copies of RNN are connected, and they communicate by passing information from one state to another. Hence, RNN with a loop signifies that it is composed of multiple copies of RNNs. The current state is computed by using the previous state value and the current input. RNNs are very popular for sentiment classification where the network classifies a piece of text into various sentiment classification tasks. The commonly used variants of RNNs are LSTM, GRU, bi-directional RNNs, and deep RNNs.

3.3.1 Long short-term memory (LSTM)

LSTM is one of the most popular variants of RNN, which possesses the capability to handle the vanishing gradient problem in standard RNN and can catch long-term dependencies. This makes them more powerful and flexible. The architecture of LSTM is shown in Fig. 10 and is comprised of forget gate f_t, input gate i_t and output gate o_t. The forget gate f_t helps in deciding which information to dump from cell at time t by taking the value of h_{t-1} (previous state information) and x_t (current input), to output a value of 0 or 1 where 0 signifies *completely dump,* and 1 signifies *completely keep.*

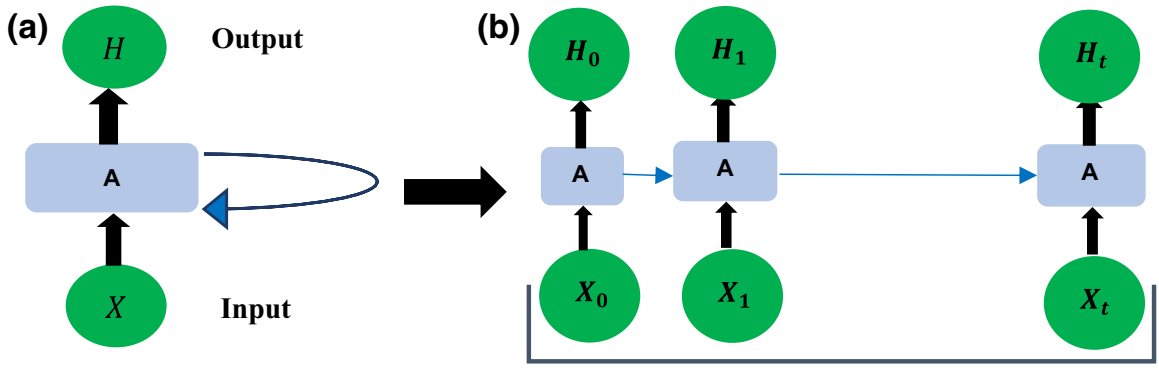


Fig. 9 a RNN with combined information from all the previous states and **b** unrolled version of RNN showing all the previous states

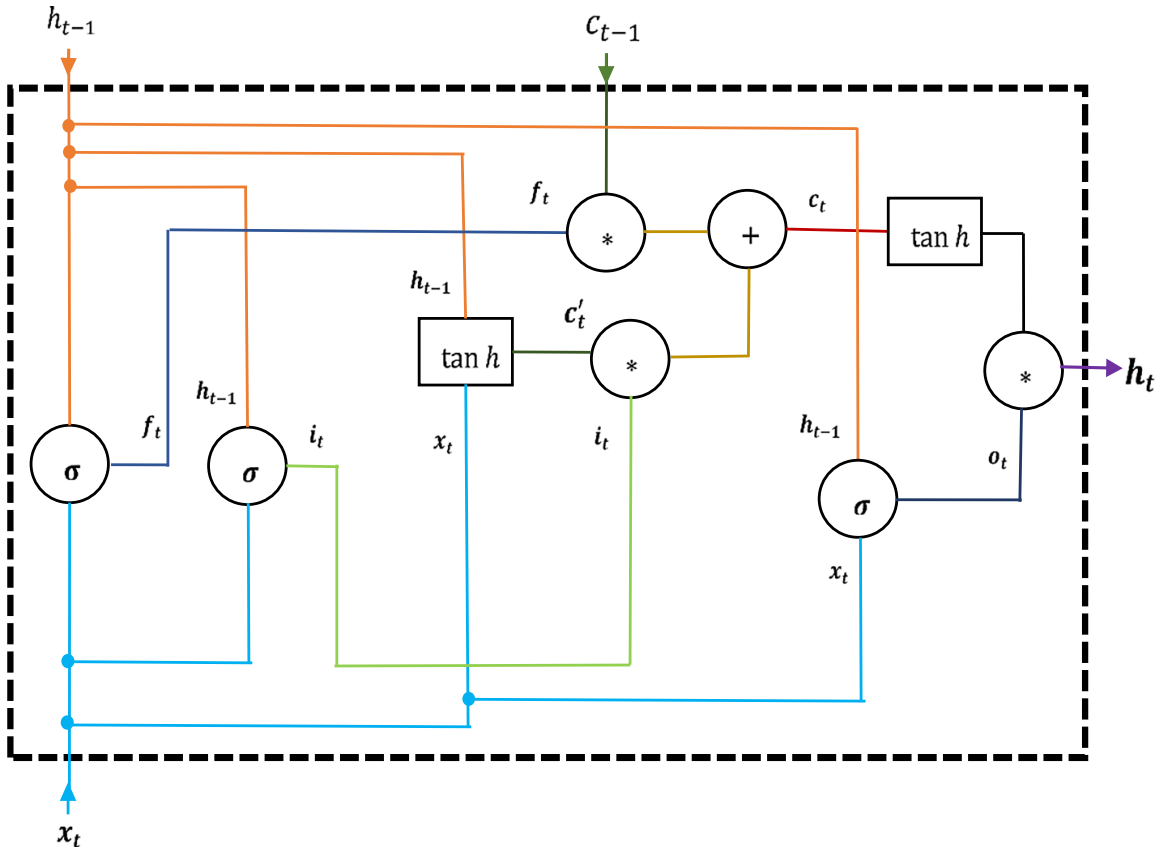


Fig. 10 Architecture of LSTM

The next step is to update the state which is done by combining these two steps: In first step, input gate i_t decides the value to be updated. In second step, *tanh* layer generates a vector with new candidate values C'_t that will be supplied to the cell's state. The old state C_{t-1} is updated to new a state C_t by multiplying it with forget gate f_t and the amount by which the state will be updated is decided by the new candidate values $i_t * C'_t$. Finally, the output gate o_t is generated by running a sigmoid layer which decides the part of the cell state that will serve as an output. Finally, for restricting the value between -1 and 1, the cell state is passed through a *tanh* layer. The resulting value is multiplied by o_t.

Researchers are also combining various deep learning techniques for sentiment classification. Huang et al. (2017) combined two popular deep learning networks called LSTM and CNN (Fig. 11) and proposed an architecture which is composed of a layer of CNN and

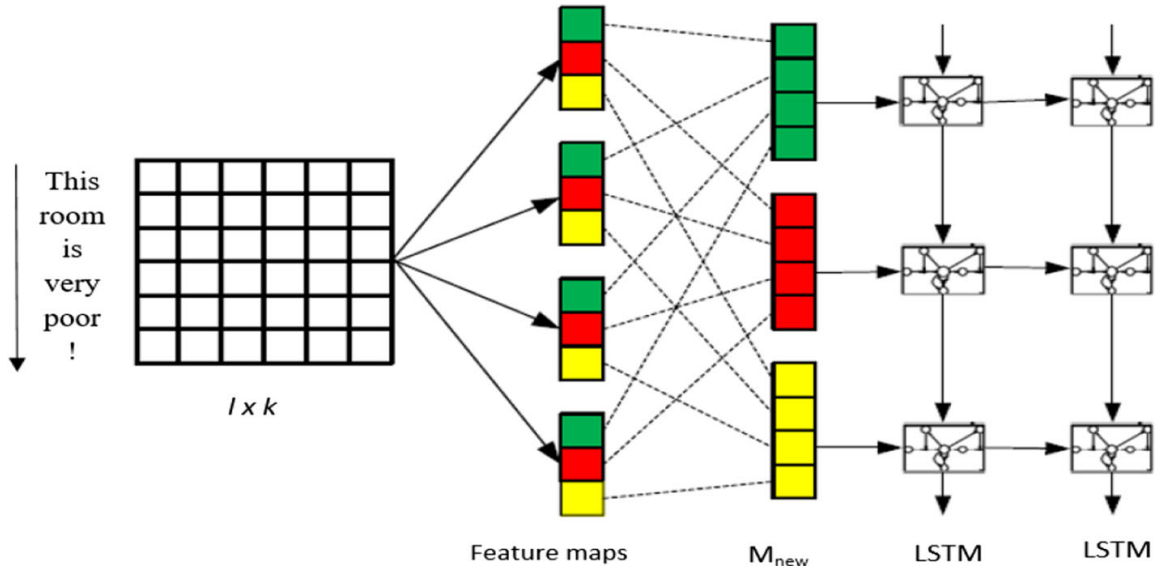


Fig. 11 Combining CNN and LSTM for sentiment classification (Huang et al. 2017)

Table 6 Comparison of work presented by Huang et al. (2017) and Hassan and Mahmood (2017b)

Refs.	Objective	Dataset	Accuracy (%)	Advantage
Huang et al. (2017)	Employed CNN to capture significant local features of the text, fed them to a two-layer LSTM model. Applied pre-trained word2vec model	Chinese Sina microblogs (binary)	87.2	The proposed model can extract context-dependent features for generating the sentence representation for sentence classification
Hassan and Mahmood (2017b)	Proposed an architecture (*ConvLstm*) which applies LSTM as a substitute of the pooling layer in CNN. Applied pre-trained word2vec model	(a) IMDB (binary) (b) SSTb (Fine-grained)	(a) 88.3 (b) 47.5	The proposed architecture can reduce the loss of local information and capture long-term dependencies

two-layer of LSTM stacked on CNN. Similarly, Hassan and Mahmood (2017b) proposed an architecture called *ConvLstm,* which again combined CNN and LSTM for classifying short texts on the top of word2vec. Table 6 shows a comparison of the work presented by both authors.

From Table 6, we observe that researchers are combining CNN + LSTM based approaches for sentence-level sentiment classification as CNN can extract the local features in the text, and LSTM can capture the long-term dependencies in sentences. Yoo et al. (2018) proposed a system called Polaris, which can analyze the sentiment trajectory of any event. Trajectory analysis can be done by classifying the contents regarding a particular event on social media according to the area where those events have occurred. The proposed model gave an F1

score of 84.1%. Chen et al. (2017b) discussed a divide-and-conquer approach for sentence-level sentiment classification. Initially, target expressions from opinionated sentences are extracted using bi-LSTM with conditional random fields (biLSTM-CRF), and sentences are classified into a non-target group, one-target group, and multi-target group. The 1d-CNNs are trained separately on each of the target groups and are used for obtaining the sentiment polarity for every type of sentence. They conducted experiments on datasets, which include movie reviews, customer product reviews, and Stanford sentiment treebank (SST-1 and SST-2) with binary and fine-grained sentiment labeled reviews along with 11 other approaches. Experimental results signifies that dividing sentences in different targets improved the performance for sentence-level sentiment analysis.

3.3.2 Gated recurrent units (GRUs)

GRUs are variants of LSTM and are considered as LSTM without an output gate. They deal with two kinds of gate: update gate and reset gate. The architecture of GRU is shown in Fig. 12. The gating mechanism of GRU is explained as follows:

(a) *Update gate* This gate defines how much information (memory) needs to be kept for the future. The input h_{t-1} denotes information from the previous $t-1$ state, x_t represents current state value, and z_t is update gate value.
(b) *Reset gate* This gate, r_t defines how much past information the network will forget.
(c) *Calculate the candidate value* The candidate value $h_t^{'}$ is calculated by taking element-wise product (denoted by, $\odot$) of set gate r_t and h_{t-1} to determine the information which needs to be removed from previous time steps, adding Ux_t where U is a parameter vector and applying *tanh* activation function on the output.
(d) *Calculate the final memory value* To get the final memory value of the current unit h_t, update gate values z_t are required to determine which information is needed from $h_t^{'}$ and which information will be required from previous value h_{t-1}.

Attention-based GRU networks are also applied to various sentiment analysis tasks like target based sentiment classification. Zhang et al. (2018b) proposed an approach to model target sentiment classification into Q&A system using Dynamic Memory Networks (DMN). The DMN consisted of four modules: Memory module consisting of input module which encodes the input sentence and uses single-layer bi-GRU to learn semantics of every word, Question module lists the questions according to different target types and Answer module shows the sentiment expressed towards the target. The memory module is comprised of multiple attention blocks and memory updating block, where multiple attention blocks consist of soft attention, attention-based GRU network, and inner attention network. The experimental results on SemEval 2014 (laptop and restaurant reviews) and Twitter dataset proved that attention-based GRU and inner attention can be used for solving the weight bias problem, thus improving target based sentiment analysis. GRUs have also been used for sentiment classification at the document-level. Tang et al. (2015a) discussed document-level sentiment classification using deep learning models. They first used CNN or LSTM to generate sentence representations from word representations and then used GRU for encoding intrinsic relations between the sentences in document representation. The experimental results showed an accuracy of 67.6% on Yelp 2015 dataset and an accuracy of 45.3% on IMDB dataset. Hence, GRU can be used to model sentence representation.

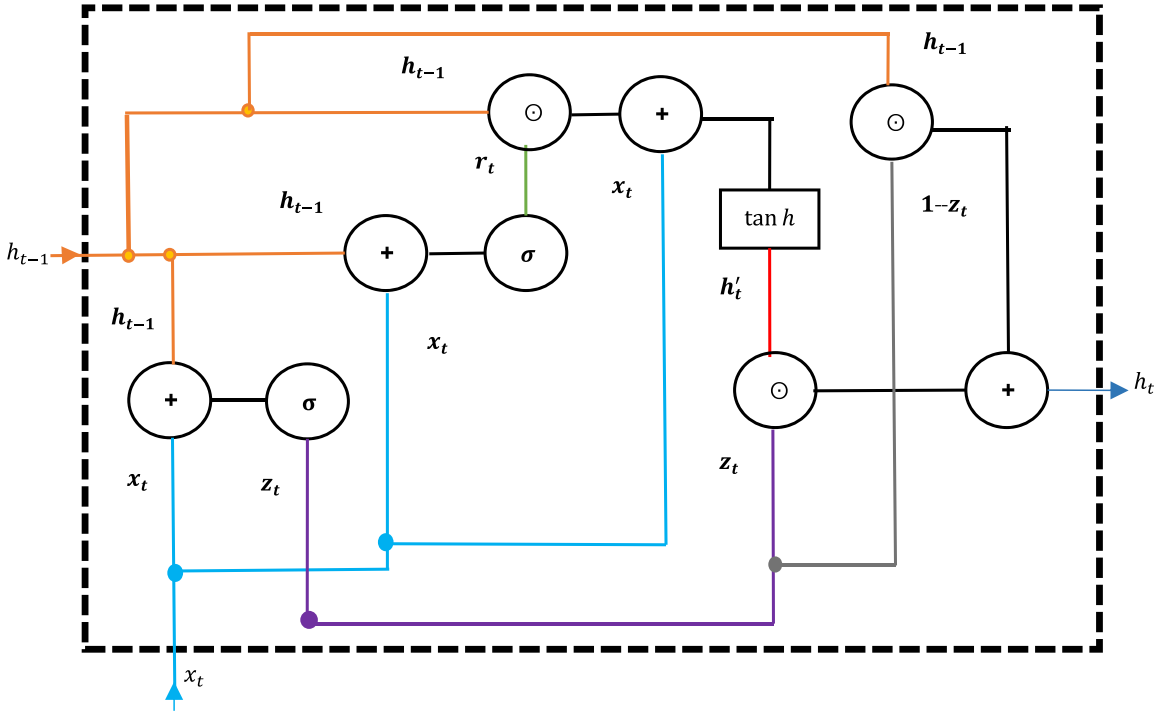


Fig. 12 Architecture of GRU

3.4 Deep belief networks (DBNs)

DBNs emanate under the type of unsupervised pre-trained networks, which also includes autoencoders and Generative Adversarial Networks. DBNs are composed of multiple layers of unsupervised networks like Restricted Boltzmann Machines (RBMs) or autoencoders as shown in Fig. 13a. Hence, various layers of RBM (Yanagimoto et al. 2013; Ruangkanokmas et al. 2016; Yuan et al. 2014) are stacked together to develop a Deep Belief Network. RBM is a stochastic neural network which consists of one layer of visible nodes $v = [V_1, V_2, V_3]$ and one layer of hidden nodes $h = [H_1, H_2]$. Each visible node is connected to hidden nodes and vice versa through an undirected connection as shown in Fig. 14a. Each hidden layer of the RBM learns the higher-level features progressively (called as pre-train phase) from the data, which are later combined for automatic feature engineering. To achieve easy learning, the restriction is made for visible nodes and hidden nodes that a visible node is not connected to other visible nodes and the same for the hidden nodes too.

A classical neural network perceptron model is shown in Fig. 14b. It is made up of three layers such as the input layer, hidden layer, and output layer. These layer consist of various nodes such as input nodes $x = [X_1, X_2, X_3]$, hidden nodes $h = [H_1, H_2, H_3, H_4]$, and output node Y and due to the use of multiple layers, it is also called a multilayer perceptron (MLP). The output of this model is computed based on input, weights, bias, and threshold parameter.

The difference between RBM and classical neural network is that an RBM has only two layers while classical neural network has three layers. RBM is an unsupervised learning (no levels) model while the classical neural network is supervised (with levels) and unsupervised both. RBM is primarily a generative model, while MLP is a discriminative model. DBNs are unsupervised networks because they are composed of unsupervised units of RBMs, while DNNs can work as a supervised or unsupervised model. DBNs are generative models, and DNNs are discriminative models. Hence, to build a DNN from a DBN, a discriminative layer

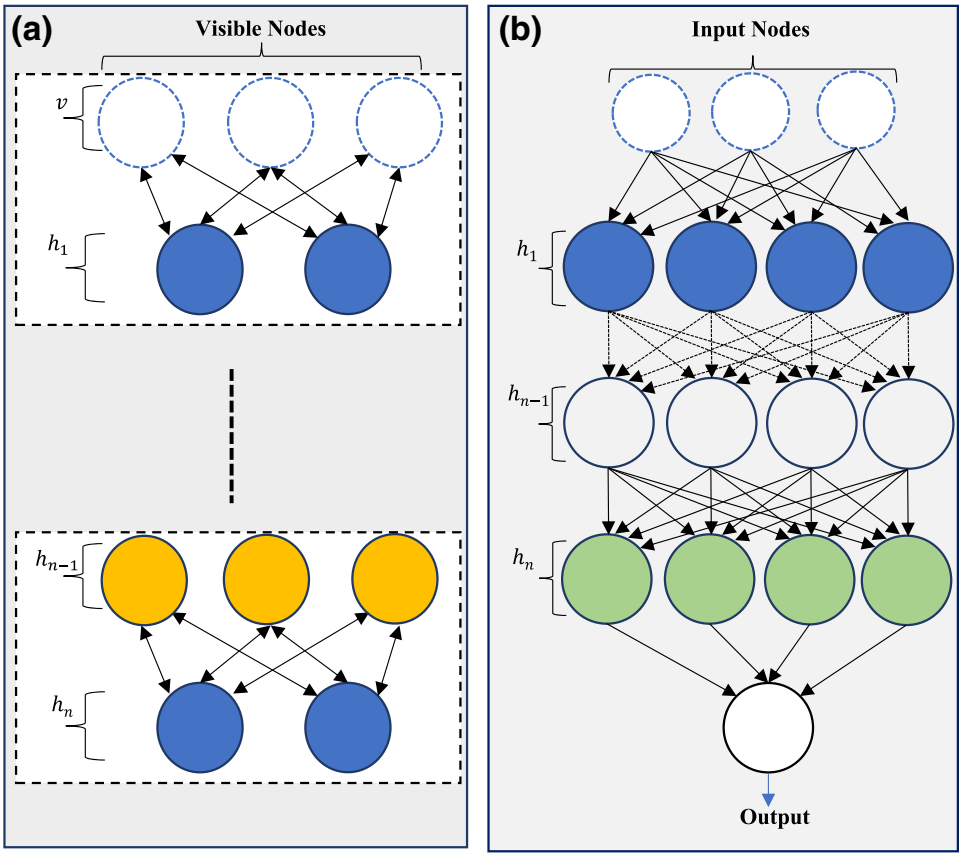


Fig. 13 **a** Deep belief network and **b** deep neural network

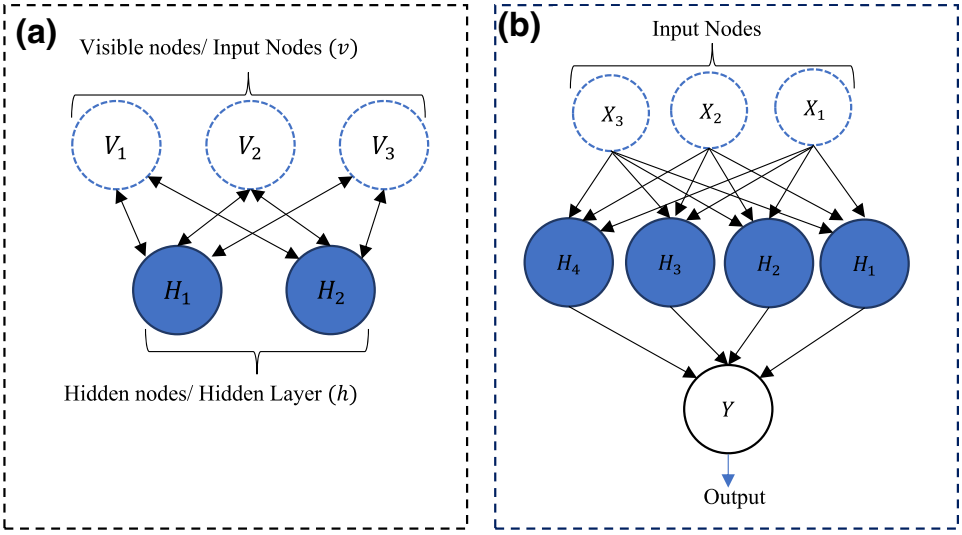


Fig. 14 **a** RBM unit and **b** classical neural network

is added on top of DBN. The key difference between DBN and DNN is in the training process as DBNs are pre-trained to reconstruct an input and then fine-tuned with backpropagation while DNN has purely supervised training with backpropagation. The pre-training of DBN is beneficial where the training set is small. Additionally, DBN supports bidirectional inter-layer communication between two different layer nodes whereas, in DNN, it is unidirectional.

Jin et al. (2017) applied DBNs with delta rule for sentiment classification on ten sentiment datasets. Delta rule uses gradient descent for fine-tuning the weights in a single layer neural

network. RBM is used to train the weights and pass them into the network through back-propagation. Experimental results show that the proposed approach (DBN with delta rule) performs better than DBN. Yanagimoto et al. (2013) developed a neural network architecture, which is composed of four layers of RBM sharing hidden layer units and visible units for estimating the similarity between the articles. The proposed approach has given good results, which shows that this architecture can be applied to the various natural language processing tasks. Ruangkanokmas et al. (2016) used DBNs with feature selection (DBNFS) for sentence classification. Initially, feature extraction is done in which pre-processing is applied to remove punctuations, stop words, and whole document is converted into feature vectors. Next, five different datasets are employed and are partitioned into an unlabeled training set, labeled training set and labeled testing set for constructing a classifier. For feature selection, Chi squared measure is used to select the most relevant features to obtain a reduced feature set. Finally, sentiment classification is done by DBNswhich outperforms the results obtained by baseline semi-supervised methods giving the best accuracy of 75.0% on electronics review dataset.

Apart from the above deep learning models, the other models which are gaining popularity in sentiment analysis are:

- *Attention-based networks* The traditional RNN approaches capture irrelevant information in the piece of information-rich text. Hence, the attention mechanism was introduced, which was inspired by the visual attention mechanism found in humans. It decides which part of the text should be focused on, rather than encoding the full sentence length. Sentiment analysis using attention-based networks is used in Yuan et al. (2018), Zhang et al. (2017, 2018b), Jiang et al. (2014), Song et al. (2018) and Yang et al. (2018).
- *Bi-directional RNN (BRNN)* A major drawback of the deep learning-based models was that they were able to learn information from previous time steps only. To overcome this issue, BRNN was introduced, which can efficiently capture the information from the future time steps for removing ambiguity and understanding the context. The module used in BRNN could be the standard RNN, LSTM, or GRU. Hence, BRNN has two types of connections: one going from left to right, in forward direction till the final time step, and other going from right to left, moving backward in time to the initial time step. Sentiment analysis using BRNN is reported in Chen et al. (2017b), Baktha and Tripathy (2017), Poria et al. (2017b) and Wang et al. (2018a).
- *Capsule networks* CNN architecture suffers from certain limitations. They were not able to model the hierarchical relationship between the local features which might misclassify objects based on their properties. The max-pooling operation in CNN results in losing certain valuable information as only the active neurons are move to the next layer. Hence, capsule networks (Sabour et al. 2017) were proposed, which can overcome these limitations. These networks consider the spatial relationships between the entities by using dynamic routing between the capsules, which is much better than max pooling operation of CNN. The dynamic routing trains the neuron vectors in capsule networks, which replaces the neuron node of the traditional neural network. The capsule networks can be trained with much less information than other neural network-based architectures. Moreover, capsule networks are getting famous in natural language processing for various text classification tasks (Du et al. 2019a; Zhao et al. 2019; Kim et al. 2019). Wang et al. (2018b) proposed RNN based capsule networks by building capsules for each sentiment category. Du et al. (2019c) proposed hybrid capsule networks for obtaining the implicit semantic information. Capsule networks are popularly used for multi-label text classification (Aly et al. 2018) and cross-domain sentiment classification (Zhang et al. 2018c; Yang et al. 2019b).

They are also applicable for aspect-level sentiment classification (Du et al. 2019b; Chen and Qian 2019). Wang et al. (2019) proposed a model in which the capsule structure can focus on each aspect category. Each capsule outputs the aspect probability and sentiment distribution on the targeted aspect. The results show that these networks can detect the words which prominently reflect the aspect category.

3.5 Application-wise comparison of deep learning approaches for sentiment analysis

This section discusses the applications of different deep learning architectures for sentiment analysis in Table 7.

3.6 Drawbacks of deep learning approaches in sentiment analysis

Although, deep learning algorithms have shown excellent outcomes and significant evolution in sentiment analysis, yet there exist some drawbacks of applying these algorithms which are stated as follows:

- Most deep learning techniques require a lot of labeled data for training to make sure that machine delivers the desired results. Hence, for sentiment analysis, we require a large corpus to properly train the deep learning model for correct prediction of the class labels. Gathering and labeling large amounts of data can be very difficult and tedious.
- Unlike traditional machine learning or lexical methods where we know what features are selected for predicting a particular sentiment, it is hard to figure out what is the actual reason for the neural network to predict a specific sentiment in a body of text just by looking at weights in different layers. This makes it difficult to get an intuition about the prediction process of the neural networks, and they behave like a "black box" to many researchers.
- Deep learning techniques like CNN requires tuning of initial parameters as a starting point. This can be seen in Stojanovski et al. (2015). Thus, the performance of the network depends upon the values of the hyperparameters of the network. Hence, deciding the optimal hyperparameter values is a challenging task.
- Due to a large number of parameters present in the deep learning models, the time taken to train them is often very large (Dufourq and Bassett 2017). Moreover, they need high performing hardware like GPUs and large RAM for better efficiency.

3.7 Merits and demerits of deep learning models

This section draws a fair comparison between different models by discussing the merits and demerits of various deep learning models in Table 8.

4 Datasets and performance measures

This section discusses the sentiment analysis dataset, evaluation measures used in sentiment analysis, execution time and performance comparison of deep learning methods.

Table 7 Application-wise comparison of deep learning approaches

Model	Refs.	Applications
CNN	Wang et al. (2016b), Tang et al. (2015b) and Do et al. (2019)	Learn the generalizable features across speakers for multimodal data (Wang et al. 2016) Capture the local semantics of n-grams of various granularities, which are proven powerful for sentiment classification (Tang et al. 2015) Identify fixed length phrases in consumer review domain (Do et al. 2019) Capture the most important n-grams feature of a sentence (Do et al. 2019)
RNN	Huang et al. (2018), Do et al. (2019) and Giachanou and Crestani (2016)	Shown great performance in Twitter domain which contains informal context and emoticons (Giachanou and Crestani 2016) Improved the performance for aspect-based sentiment analysis (Do et al. 2019) Not able to process long text due to vanishing gradient problem (Huang et al. 2018)
LSTM	Huang et al. (2018), Abdi et al. (2019) and Kraus and Feuerriegel (2019)	For text-based reviews, they are able to represent the semantic structure of a document (Kraus and Feuerriegel 2019) and model the semantic relationship in words (Abdi et al. 2019). Retain the information and order about each word in the text (Abdi et al. 2019) Able to model long text in document-level sentiment classification (Huang et al. 2018)
GRU	Zhang et al. (2018b)	BiGRU capture hidden semantics of every word in a sentence (Zhang et al. 2018c) Less computationally expensive than LSTM (Zhang et al. 2018c) Efficiently model the interaction between targeted words (Zhang et al. 2018c)
Attention	Chen (2017) and Cheng et al. (2017c)	Temporal attention + LSTM improves the sentiment prediction despite the presence of noisy acoustic and visual modalities (Chen 2017) Hierarchical attention can model the aspect information of the text, which further helps in capturing the sentiment information (Cheng et al. 2017c)
Rec NN	Do et al. (2019) and Tai et al. (2015)	Can represent multiple aspects within the text (Do et al. 2019) Learn new words and capture tree-structure of sentences (Tai et al. 2015)
DBN	Sun et al. (2016)	RBM units inside the DBN can learn the hidden structures in the input data for high-level feature representation (Sun et al. 2016)

Table 8 Merits and demerits of different deep learning models

Model	Merits	Demerits
CNN (Nogueira et al. 2014; Lee et al. 2018)	Faster and computationally less expensive as compared to RNN, LSTM, and GRU networks Fewer parameters have to be trained at the time of back propagation which makes the training significantly faster CNN combined with other networks such as LSTM has shown high accuracy Has the ability to capture relevant features from any part of the word For text, CNN models can identify relatively long sentiment/emotion phrases	Fails to preserve long-term dependency CNN ignores the long-distance dependency features that reflect syntactic and semantic information which is crucial for understanding sentences in sentiment analysis
RecNN (Socher et al. 2013; Singhal and Bhattacharyya 2016)	Powerful for learning tree-like hierarchical structure. Hence, it can be good for NLP tasks, since grammar follows a tree-like structure Have the ability to capture negation in the input sentence which is a crucial step for sentiment classification	Training can be hard as the structure changes for every training sample Model is still in its infancy stage In the case of informal data (like Tweets) which do not follow any grammatical rules, the efficiency of the model is highly affected
RNN (Hassan and Mahmood 2018)	Pays attention to the sequence of words (sequential data) in the data, so it performs better than CNN Able to store past computations Keeps the number of parameters less by weight sharing Can capture the long-distance dependency features that reflect syntactic and semantic information which are ignored by CNN Helps to preserve ordering of words in an input sequence	Not able to process very long sequences May suffer from the vanishing gradient and exploding gradients problem which makes it unfavorable for NLP tasks like sentiment analysis

Table 8 continued

Model	Merits	Demerits
LSTM (Huang et al. 2018; Peng et al. 2018; Singhal and Bhattacharyya 2016)	Can selectively forget and remember things which makes it much better for classifying sentiments Able to prevent the vanishing and exploding gradient problem in an RNN by remembering inputs from previous steps Performs much better than CNNs because of its ability to take the sequence of text in the account for predicting the sentiment For text, they can help in encoding semantics of a sentence and their representation Can be used to extract sequential information in aspect target sequence for aspect-based sentiment analysis	Computationally expensive to apply back propagation after calculating the output each time As each weight has to be trained, it can also make it harder to find an optimal solution. This also makes this network considerably slower Output produced at each time step (for every input word) needs to be reconciled to a whole sentence to get the label for the entire sentence
GRU (Verma et al. 2018; Rana 2016)	Has a fewer number of gates, hence is faster and computationally less expensive Solves the problem of vanishing and exploding gradient that is present in RNN Less complex structure than LSTM Can be useful for capturing interdependencies that exist between sentences of review document in document-level sentiment classification Have the ability to recognize emotion from a noisy speech	Does not have a memory unit, hence exposes the full hidden content without control On larger texts, LSTMs perform better than GRU because of their ability to retain memory for a longer time
DBN (Yanagimoto et al. 2013; Jin et al. 2017)	Able to learn useful information like a huge scale dimension of vocabulary from a corpus using several hidden layers	Unable to remember the previously trained task, unlike LSTM due to lack of memory Time-consuming, computationally expensive for large applications

4.1 Sentiment datasets

In order to test the performance of sentiment analysis algorithms, dataset plays a significant role. The performance is calculated in terms of prediction accuracy or F1 score and the bold letter text in Tables 9, 10 and 11 indicates the highest accuracy and model on the respective dataset. We have identified some of the popular datasets for sentiment analysis. Table 9 describes the datasets applied for sentiment analysis by describing the main features of a dataset, the deep learning models applied to them, and accuracy (or F1 score) achieved on it.

From Table 9, we conclude that some of the popular datasets which are used by various researchers in the area of sentiment analysis with deep learning models are as follows:

- *Stanford large movie review (IMDB)* (Maas et al. 2011) is a publicly available dataset consisting of 50,000 binary labeled movie reviews partitioned evenly for negative and positive reviews. A hybrid approach of RNN + CNN has shown excellent results on this dataset.
- *Yelp dataset* (Yelp Dataset 2014) consists of restaurant review labeled on a scale of 1–5 derived from the Yelp Dataset Challenge. RNN and its variants like Bi-RNN and LSTM are popularly applied to this dataset.
- *Stanford sentiment treebank (SSTb)* (Socher et al. 2013)is also a publicly available dataset containing 11,855 movie reviews from *rottentomatoes.com* website. The SSTb dataset includes five classes for classification (fine-grained sentiment classification). A hybrid approach of RNN + CNN has shown the highest results on this dataset.
- *Amazon review dataset* (Blitzer et al. 2007) is composed of product reviews from Amazon on four different types of products: books, DVDs, electronics, and kitchen appliances. Review with ratings greater than three is classified as positive and review with ratings less than three is classified as negative. A combined approach of CNN + BiLSTM + Attention has shown the highest accuracy of 87.76%.
- *CMU-MOSI dataset* (Zadeh et al. 2016) is a popular multimodal dataset which consists of 2199 opinionated utterances from 93 videos on different topics crawled from YouTube. Attention-based LSTM with dynamic fusion is the popular approach for this dataset.
- *MOUD dataset* (P'erez-Rosas et al. 2013) is another popular multimodal dataset which contains 79 videos about product reviews in Spanish. Google translate API2 is used for conversion from Spanish to English transcripts. Bi-LSTM model has shown the highest accuracy with 71.1%.
- *Getty images dataset* (You et al. 2016) consists of 588,221 labeled data that contains both images and text. Tree LSTM with Attention has given the highest accuracy on this dataset.
- *Twitter dataset* (You et al. 2016) consists of 220,000 tweets that contain both images and text. Similarly, Tree LSTM with Attention has given the highest accuracy on this dataset.
- *Twitter image dataset* (You et al. 2015) consists of 1269 image tweets labeled by five Amazon Mechanical Turk (AMT) workers. Fine-tuned CNN (GoogleNet) has given maximum accuracy on this dataset.

Hence, sentiment analysis datasets mainly comprise: Tweets, debates, messages, comments, blogs, posts, hashtags, consumer reviews on Google Play, multi-domain review about books, electronics, and kitchen appliances, reviews about restaurant, online products, hotels, and places (or locations), star ratings about products, emoticons, and images. The highest accuracy obtained on IMDB dataset is 93.2% by using CNN + RNN, and on Amazon multi-domain dataset highest accuracy obtained is around 87% by using a hybrid approach of CNN + bi-LSTM + Attention model. Hence, the RNN and its variants like LSTM, bi-RNN, bi-LSTM are giving high accuracy rates for sentiment analysis tasks, which justifies the pop-

Table 9 Description of the dataset for SA

Name of the dataset	Main feature	Deep learning models	Accuracy or F1 score (%)
Yelp 2013 (Yelp Dataset 2014)	Yelp 2013 dataset consists of restaurant review labeled on a scale of 1–5 derived from Yelp Dataset Challenge	**LSTM + KNN** (Zhou et al. 2018)	**70** (Average F1 score)
		BiLSTM (Chen et al. 2019)	56.3
		GRU (Li et al. 2019)	68.04
		Sliced RNN (Li et al. 2019)	66.02
		CNN (Tang et al. 2015)	59.6
		Bi-RNN + Attention (Zhang and Chow 2019)	**70.6**
Yelp 2014 (Yelp Dataset 2014)	Yelp 2014 dataset consists of restaurant review labeled on a scale of 1–5 derived from Yelp Dataset Challenge	GRU (Li et al. 2019)	70.90
		Sliced RNN (Li et al. 2019)	**71.60**
		CNN (Du et al. 2019d)	61.4
		Convolution + Recurrent (Du et al. 2019)	71.5
		LSTM + GRNN (Du et al. 2019)	67.1
		LSTM + KNN (Zhou et al. 2018)	69 (Average F1 score)
IMDB (Maas et al. 2011)	IMDB consists of movie reviews labeled on a scale of 1–10	LSTM + KNN (Zhou et al. 2018)	52.2 (Average F1 score)
		CNN (Lee et al. 2018)	89.57
		Joint RNN and CNN framework (Hassan and Mahmood 2018)	**93.2**
		Bi-RNN + Attention (Zhang and Chow 2019)	56.9
		Convolution + Attention (Du et al. 2019)	50.2
SSTb (Stanford Sentiment Treebank) (Socher et al. 2013)	SSTb contains labeled reviews in 5 classes for multi-class classification	LSTM (Donnelly and Roegiest 2019)	88.52
		RNN (Socher et al. 2013)	79.0
		RNTN (Socher et al. 2013)	80.7
		Bi-GRU (Baktha and Tripathy 2017)	43.81
		Joint RNN and CNN framework (Hassan and Mahmood 2018)	**89.2**

Table 9 continued

Name of the dataset	Main feature	Deep learning models	Accuracy or F1 score (%)
Amazon (Blitzer et al. 2007)	Amazon dataset contains reviews fromBooks, DVD, Electronics, and Kitchen domain from 25 products	Hierarchial attention network (HAN) (Manshu and Bing 2019)	81.07
		CNN (Manshu and Bing 2019)	81.98
		3 layer CNN + BiLSTM + Attention (Manshu and Bing 2019)	**87.76**
		Attention + BiLSTM (Yuan et al. 2018)	87.69
CMU-MOSI (Multimodal Corpus of Sentiment Intensity dataset) (Zadeh et al. 2016)	MOSI dataset consists of 2199 opinionated utterances from 93 videos on different topics crawled from YouTube	LSTM + Temporal Attention + Word level fusion (Chen 2017)	75.7
		LSTM + Hierarchical Fusion (Poria et al. 2017b)	80.30
		Attention based LSTM + Dynamic feature fusion (Poria et al. 2017c)	**81.3**
		Select-Additive Learning + CNN (Wang et al. 2016)	73
		Temporal Attention model (Yu et al. 2017)	75.1
MOUD (P'erez-Rosas et al. 2013)	It contains 79 videos about product reviews in Spanish. Google translate API2 is used to for conversion from Spanish to English transcripts	LSTM + Hierarchical Fusion (Poria et al. 2017)	68.11
		Attention based CNN and RNN + audio feature fusion (Luo et al. 2019)	68.74
		Bi-LSTM (Li and Xu 2019)	**71.1**

Table 9 continued

Name of the dataset	Main feature	Deep learning models	Accuracy or F1 score (%)
Getty Images (You et al. 2016)	It consists of 588,221 labeled dataset that contains both images and text	CNN + Distributed paragraph vector (You et al. 2016) CNN + LSTM + Attention + late fusion (Huang et al. 2019) **Tree LSTM + Attention** (Trofimova et al. 2016)	80.0 86.9 **90.2**
Twitter (You et al. 2016)	It consists of 220,000 tweets that contains both images and text	CNN + Distributed paragraph vector (You et al. 2016) CNN + LSTM + Attention + late fusion (Huang et al. 2019) **Tree LSTM + Attention** (Trofimova et al. 2016)	80.9 76.3 **96.4**
Twitter image dataset (You et al. 2015)	It consists of 1269 image tweetslabelled by five Amazon Mechanical Turk (AMT) workers	CNN (You et al. 2015) Fine-tuned CNN +oversampling (Campos et al. 2015) Fine-tuned CNN (Alexnet) (Wang et al. 2016c) **Fine-tuned CNN (GoogleNet)** (Islam and Zhang 2016) CNN +Attention (Song et al. 2018)	78.3 83.0 83.8 **86.1** 85.1

Table 10 Comparison of execution time of different deep learning algorithms

Refs.	Sentiment analysis task	Dataset (#Training sample instances)	Deep learning architecture	Execution or training time (s)	Platform	Highest accuracy (%)
Zhao et al. (2017)	Sentence classification	Review from Amazon on digital cameras, cell phones and laptops (600)	CNN **LSTM**	1800 18,000 (Execution time)	Nvidia GTX 980Ti GPU	87.9
Li et al. (2019)	Sentence classification	Yelp 2015 (808,052)	GRU **Bi-GRU** Sliced RNN Bi- Sliced RNN	1609 3176 218 440 (Training time)	Keras, NVIDIA GTX 1080Ti GPU	73.36
Li et al. (2017b)	Sentence classification	Online debates (24352), Restaurants (2614) and laptop reviews (5485) from SemEval 2014 and 2015	CNN LSTM MemNet **AttNet**	CNN:5 LSTM:200 MemNet:150 AttNet:200 (Training time)	TITAN X GPU	(Avg F-score) Debates: 52.23 Tweets: 35.34 Review: 55.93
Tay et al. (2017)	Aspect-based sentiment analysis	Customer reviews for laptop (1813), restaurants (3102), SemEval 2014 (3587), Tweets from SemEval 2016 (2771), Online Debates (24564)	**Memory NN** LSTM Attention- LSTM	Memory NN:6 LSTM:9 Attention- LSTM: 12 (Execution time)	NVIDIA GTX 1070 GPU	(Overall F1 score) 69.2
Yuan et al. (2018)	Multi-domain sentiment classification	Amazon multi-domain dataset (Amazon) (Blitzer et al. 2007) (1400) Sanders Twitter Sentiment Dataset (Sanders)	RNN GRU LSTM **BiLSTM + attention**	RNN:71 GRU:246 LSTM:310 LSTM with peephole connection:411 (Training time)	TensorFlow, NVIDIA K80 GPU	(Avg accuracy) Amazon: 87.69 Sanders: 86.32

Table 11 Performance comparison of deep learning algorithms

Refs.	Dataset	Technique/Method	Accuracy (%)	# Instances in training set
Baktha and Tripathy (2017)	Amazon health product reviews	Vanilla RNN LSTM **GRU** Bi-Vanilla RNN Bi-LSTM Bi-GRU	57.30 78.10 **83.90** 58.00 79.20 81.10	Amazon: 8000
Xu et al. (2011)	Amazon mobile phone reviews	**Multi-class SVM** CRF without interdependencies CRF with interdependencies	**61.38** 60.04 66.17	–
Rain (2013)	Amazon books, media, and kindle product reviews	Decision Tree **Naïve Bayes**	79.84 (for books) **84 (for kindle)**	–
Shaikh and Deshpande (2016)	Amazon books, camera, music product reviews	Naïve Bayes (Books) Naïve Bayes (Camera) Naïve Bayes (Music)	80 (multiword level feature) 80 (single word level feature) 80 (single word level feature)	260
Al-Smadi et al. (2017)	Arabic Hotels' Reviews (Al-Smadi et al. 2016; Pontiki 2016)	For aspect-based sentiment analysis tasks Deep RNN **SVM**	For Sentiment Polarity Identification RNN:87 **SVM:95.4**	19,226

ularity of RNN for modeling sequential data. Moreover, the combination of CNN + RNN is used for images and video-related data for visual sentiment analysis.

Most of the researchers experimented by creating their own corpus. Yang et al. (2018) created their Chinese corpus as no dataset was publicly available for target-dependent sentiment classification. Shi et al. (2017) needed user's information for user-based features. So, they crawled the weibos from Sina Weibo and labeled the weibos as positive, negative, or neutral. Due to the unavailability of publicly accessible Chinese corpus, Li et al. (2014) built a labeled parse tree corpus called "Chinese Sentiment Treebank" which is a collection of reviews about 2270 movies from a popular movie review website and made it publicly available. Hence, new datasets are needed to train the model better with a large number of records. Earlier the size of the dataset was limited to thousands of records. But now, large datasets are being formed which contain millions of records as training examples.

4.2 Evaluation measures

Most of the state-of-the-arts for sentiment analysis uses accuracy (Hassan and Mahmood 2018; Lee et al. 2018; Zhang and Chow 2019), F1 score (Xiong et al. 2018b; Dragoni and Petrucci 2017; Tay et al. 2017; Zhou et al. 2018), precision (Dragoni and Petrucci 2017; Tay

et al. 2017; Zhou et al. 2018), and recall (Dragoni and Petrucci 2017; Tay et al. 2017; Zhou et al. 2018) as performance measuring parameters. These measures are defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F1 score} = \frac{2 * Precision * Recall}{Precison + Recall} \tag{4}$$

where TP is the true positive, FP is the false positive, TN is the True Negative, and FN is the False Negative.

Additionally, for proper interpretation of results, there are other parameters which have been used, such as Mean Square Error, Ranking loss, Macro-averaged Mean absolute error, and RandIndex.

$$\text{Mean Square Error (MSE)} = \frac{\sum_{i=1}^{N} (Actual_i - Predicted_i)^2}{N} \tag{5}$$

MSE can be used to measure the divergence between predicted sentiment labels and actual sentiment labels (Verma et al. 2018). The sentiment models are trained by minimizing the MSE (Jiang et al. 2014; Wang et al. 2018c).

Similarly, Ranking loss measures the average distance between the true sentiment value and the predicted sentiment value for m *sentiment* classes with n number of test samples (Moghaddam and Ester 2010). It is defined as in Eq. (6) below:

$$\text{Ranking loss} = \sum_{i=1}^{n} \frac{\left| y_i - \hat{y}_i \right|}{m * n} \tag{6}$$

Macro-averaged Mean absolute error (MAE^M) is used for handling the imbalanced datasets (Marcheggiani and Oscar 2014; Baccianella et al. 2009).

$$\text{MAE}^M(y, \hat{y}) = \frac{1}{m} \sum_{j=1}^{k} \frac{1}{|\boldsymbol{y}_j|} \sum_{y_i \in \boldsymbol{y}_j} \left| y_i - \hat{y}_i \right| \tag{7}$$

where y is true sentiment value, $\hat{y}$ is predicted sentiment value, m is the number of sentiment labels, $\boldsymbol{y}_j$ is the subset of review corpus whose true label is j.

RandIndex is generally used for clustering problems for determining the similarity between two data clustering (Zhao et al. 2014; Xu et al. 2017; Jaffali et al. 2014). The value ranges between 0 (data clustering do not have a common pair of points) and 1 (data clustering is exactly the same). It is defined as in Eq. (8):

$$\text{RandIndex}\,(C_i, C_m) = \frac{2(x + y)}{k * (k - 1)} \tag{8}$$

where C_i *and* C_m represents the clusters produced by model i and by manual annotation m, k denotes count of detected aspects, x represents number of pairs assigned to the same cluster, and y signifies the number of pairs assigned to different clusters.

4.3 Execution time comparison of deep learning algorithms

This section compares the work proposed by various researchers by discussing the execution time or training time taken by different algorithms, as discussed in Table 10.

The architecture giving the highest accuracy is marked in Bold. As seen in Table 10, LSTM and its variants (BiLSTM, GRU) require high training and execution time as compared to other deep networks like CNN and Memory networks. However, this can be compensated by high accuracy achieved by the former. Hence, a trade-off between the execution time (or training time) with overall accuracy achieved by the network becomes a crucial task for any deep learning architecture.

4.4 Performance comparison of deep learning algorithms

This section discusses some more real text examples about Google Play reviews, Amazon reviews, and Hotel reviews to compare the performance of different deep learning algorithms on various datasets with other popular machine learning algorithms like NB, SVM to see where deep learning performs correct classifications and where it performs poorly. Table 11 explains the different techniques or methods applied to sentiment analysis.

As we see, on Amazon Review dataset, the best performance was shown by GRU with 83.90% accuracy. Apart from showing great performance in many dataset and tasks, deep learning techniques still performs poorly in many cases which can be seen in Al-Smadi et al. (2017), as we observe for aspect-based sentiment analysis, SVM has outperformed RNN in Arabic Hotel Review dataset which justifies the fact that some of the machine learning algorithms are outperforming deep learning algorithms. This may be due to the ability of SVM to extract rich hand-crafted feature set for training the model and its efficiency in binary classification tasks. Hence, determining optimal deep learning architecture is a challenging task because the performance depends on the size of dataset, type of domain or area, choosing correct hyperparameters like no. of filters, no. of hidden units, filter dimension, etc.

5 Conclusion and future trend

In this article, we reviewed the most noteworthy work on sentiment analysis using deep learning-based architectures. Firstly, we introduced sentiments and its different types, followed by their application or importance for sentiment analysis. Then, we presented a taxonomy of sentiment analysis, which included Handcrafted features based approaches, and Machine learned features based approaches along with the year-wise analysis from 2011 to 2019 for sentiment analysis including deep learning approaches only. We discussed popular deep learning models which include CNNs, Rec NNs, RNNs, LSTM, GRU, and Deep Belief Networks along with their architecture, and the important and famous work using these architectures in sentiment analysis. We have also provided a brief overview of Attention-based networks, Bi-directional RNNs, and capsule networks, which have recently gained attention of researchers. In addition, we have identified sentiment analysis tasks and discussed the deep learning models applied to them. We concluded that LSTM had given better results compared to other deep learning models. We found that different languages on which sentiment analysis is applied are: English, Chinese, Arabic, Japanese, Vietnamese, Thai, Tibetan, Persian, Spanish, and Punjabiwhereas for bilingual sentiment classification Chinese and English along with Korean and English becomes the desired approach. Finally,

we have explored primary sentiment analysis dataset, main features of the dataset, deep learning model applied to them, and the accuracy (or F1 score) obtained from dataset. We also discussed the significance of creating new datasets by various researchers, drawbacks of applying deep learning in sentiment analysis, and the merits and demerits of numerous deep learning models. We reviewed around 200 articles, and based on the detailed scrutiny of different deep learning based approaches and their state-of-the-art performances discussed in this paper; we can say that sentiment analysis using deep learning approaches is a promising research area.

We have identified and presented some datasets corresponding to various sentiment analysis tasks. The major challenge faced by many researchers is the lack of proper training datasets for different sentiment analysis tasks. Moreover, deep learning methods require huge dataset for training the model. Some researchers faced problems due to the limited size of the sentiment dataset for text retrieval tasks as the dataset contained only sentiment tags but lacked the topic tags. Fine-grained sentiment analysis can also be improved by using a large number of training examples. A strong drift is seen in transfer learning based approaches for sentiment analysis. Transfer learning can be used for transferring information from one domain to another, for building robust datasets.

The field of sentiment analysis will be highly beneficial for many domains in future such as implicit sentiment detection, spam detection, temporal tagging, and stance detection. It can be applied in the medical domain to evaluate the mental health of the patients. It may be useful by many organizations for security purposes by screening the employees to validate their integrity. Moreover, emotion and genre of a movie can also be predicted by just watching its trailer. Nowadays, people are focusing more on multimodal data for analyzing sentiments. Thus, apart from incorporating star rating and user rating in the form of text, one can opt for an exhaustive rating of products with the help of multimodal data, where the reviews of a product can be incorporated using various modalities like voice, image or emoticon.

References

Abbasi A, Chen H, Salem A (2008) Sentiment analysis in multiple languages: feature selection for opinion classification in web forums. ACM Trans Inf Syst 26(3):12:1–12:34

Abdi A, Shamsuddin SM, Hasan S (2018) Machine learning-based multi-documents sentiment-oriented summarization using linguistic treatment. Expert Syst Appl 109:66–85

Abdi A, Mariyam S, Hasan S, Piran J (2019) Deep learning-based sentiment classification of evaluative text based on Multi-feature fusion. Inf Process Manag 56(4):1245–1259

Agarwal A, Yadav A, Vishwakarma DK (2019) Multimodal sentiment analysis via RNN variants. In IEEE international conference on big data, cloud computing, data science and engineering (BCD), pp 19–23

Al-Smadi M, Al-Ayyoub M, Al-Sarhan H, Jararwell Y (2016) Using aspect-based sentiment analysis to evaluate Arabic news affect on readers. In: IEEE/ACM 8th international conference on utility and cloud computing, vol 22, no 5, pp 630–649

Al-Smadi M, Qawasmeh O, Al-Ayyoub M, Jararweh Y, Gupta B (2017) Deep recurrent neural network vs. support vector machine for aspect-based sentiment analysis of Arabic hotels' reviews. J Comput Sci 27:386

Al-Smadi M, Talafha B, Al-Ayyoub M, Jararweh Y (2018) Using long short-term memory deep neural networks for aspect-based sentiment analysis of Arabic reviews. Int J Mach Learn Cybern. https://doi.org/10.1007/s13042-018-0799-4

Aly R, Remus S, Biemann C (2018) Hierarchical multi-label classification of text with capsule networks. In: Proceedings of the 35th international conference on machine learning, Sweden

Arun K, Srinagesh A, Ramesh M (2017) Twitter sentiment analysis on demonetization tweets in India using R language. Int J Comput Eng Res Trends 4(6):252–258

Azeez J, Aravindhar DJ (2015) Hybrid approach to crime prediction using deep learning. In: International conference on advances in computing, communications and informatics (ICACCI), pp 1701–1710

Baccianella S, Esuli A, Sebastiani F (2009) Multi-facet rating of product reviews. In: European conference on information retrieval. Springer, Berlin, pp 461–472
Baccianella S, Esuli A, Sebastiani F (2010) SentiwordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of the seventh conference on international language resources and evaluation (LREC'10), pp 2200–2204
Baktha K, Tripathy BK (2017) Investigation of recurrent neural networks in the field of sentiment analysis. In: Proceedings of the 2017 IEEE international conference on communication and signal processing, ICCSP 2017, pp 2047–2050
Balazs JA, Velásquez JD (2016) Opinion mining and information fusion: a survey. Inf Fusion 27:95–110
Baly R, Hajj H, Habash N, Shaban KB, El-Hajj W (2017) A sentiment treebank and morphologically enriched recursive deep models for effective sentiment analysis in Arabic. ACM Trans Asian Low-Resour Lang Inf Process 16(4):23
Beigi G, Maciejewski R, Liu H (2016) an overview of sentiment analysis in social media and its applications in disaster relief. Stud Comput Intell 639:313–340
Bhardwaj A, Narayan Y, Vanraj P, Dutta M (2015) Sentiment analysis for indian stock market prediction using sensex and nifty. In: Procedia computer science, vol 70, pp 85–91
Blitzer J, Dredze M, Pereira F (2007) Biographies, bollywood, boom-boxes and blenders: domain adaptation for sentiment classification. Annu Meet Comput Linguist 45(1):440
Bollen J, Mao H, Zeng X (2011) Twitter mood predicts the stock market. J Comput Sci 2(1):1–8
Borth D, Ji R, Chen T, Breuel T, Chang S-F (2013) Large-scale visual sentiment ontology and detectors using adjective noun pairs. In: Proceedings of 21st ACM international conference on multimedia—MM'13, pp 223–232
Brody S, Elhadad N (2010) An unsupervised aspect-sentiment model for online reviews. In: The 2010 annual conference of the North American chapter of the Association for Computational Linguistics, pp 804–812
Cambria E (2016) Affective computing and sentiment analysis. IEEE Intell Syst 31(2):102–107
Campos V, Salvador A, Jou B, Giró-i-nieto X (2015) Diving deep into sentiment: understanding fine-tuned CNNs for visual sentiment prediction. In: Proceedings of the 1st international workshop on affect and sentiment in multimedia. ACM, pp 57–62
Cao K, Rei M (2016) A joint model for word embedding and word morphology. In: Proceedings of the 1st workshop on representation learning for NLP, pp 18–26
Chachra A, Mehndiratta P, Gupta M (2017) Sentiment analysis of text using deep convolution neural networks. In: Tenth international conference on contemporary computing, pp 1–6
Chandankhede C, Devle P, Waskar A, Chopdekar N, Patil S (2016) ISAR: implicit sentiment analysis of user reviews. In: International conference on computing, analytics and security trends (CAST), College of Engineering Pune, India, pp 357–361
Chaturvedi I, Cambria E, Welsch RE, Herrera F (2018) Distinguishing between facts and opinions for sentiment analysis: survey and challenges. Inf Fusion 44:65–77
Chen M (2017) Multimodal sentiment analysis with word-level fusion and reinforcement learning. In: Proceedings of the 19th ACM international conference on multimodal interaction. ACM, pp 163–171
Chen Z, Qian T (2019) Transfer capsule network for aspect level sentiment classification. In: Proceedings oft he 57th annual meeting of the Association for Computational Linguistics, pp 547–556
Chen X, Wang Y, Liu Q (2017a) Visual and textual sentiment analysis using deep fusion convolutional neural networks. arXiv preprint arXiv:1711.07798
Chen T, Xu R, He Y, Wang X (2017b) Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. Expert Syst Appl 72:221–230
Cheng J, Zhao S, Zhang J, King I, Zhang X, Wang H (2017c) Aspect-level sentiment classification with HEAT (hierarchical attention) network. In: Proceedings of the 2017 ACM on conference on information and knowledge management, pp 97–106
Chen F, Ji R, Su J, Cao D, Gao Y (2018) Predicting microblog sentiments via weakly supervised multimodal deep learning. IEEE Trans Multimed 20(4):997–1007
Chen B et al (2019) Embedding logic rules into recurrent neural networks. IEEE Access 7:14938–14946
Collobert R, Weston J (2008) A unified architecture for natural language processing: deep neural networks with multitask learning. In: Proceedings of the 25th international conference on machine learning. ACM, pp 160–167
Day MY, Da Lin Y (2017) Deep learning for sentiment analysis on google play consumer review. In: Proceedings of 2017 IEEE international conference on information reuse and integration, IRI, pp 382–388
Do HH, Prasad PWC, Maag A, Alsadoon A (2019) Deep learning for aspect-based sentiment analysis: a comparative review. Expert Syst Appl 118:272–299
Donnelly J, Roegiest A (2019) On interpretability and feature representations: an analysis of the sentiment neuron. In: European conference on information retrieval. Springer, Cham, pp 795–802

Dragoni M, Petrucci G (2017) A neural word embeddings approach for multi-domain sentiment analysis. IEEE Trans Affect Comput 8(4):457–470
Dragoni M, Tettamanzi AGB, Pereira CDC (2016) DRANZIERA: an evaluation protocol for multi-domain opinion mining. In: Tenth international conference on language resources and evaluation, LREC, pp 267–272
Du C et al (2019a) Investigating capsule network and semantic feature on hyperplanes for text classification. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing, pp 456–465
Du C et al (2019b) Capsule network with interactive attention for aspect-level sentiment classification. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing, pp 5492–5501
Du Y, Zhao X, He M, Guo W (2019c) A novel capsule based hybrid neural network for sentiment classification. IEEE Access 7:39321–39328
Du J, Gui L, He Y, Xu R, Wang X (2019d) Convolution-based neural attention with applications to sentiment classification. IEEE Access 7:2169–3536
Dufourq E, Bassett BA (2017) EDEN: evolutionary deep networks for efficient machine learning. In: IEEE pattern recognition association of South Africa and robotics and mechatronics international conference, pp 110–115
Facebook Statistics (2019). https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/
Fernández-Gavilanes M, Alvarez-López T, Juncal-Martínez J, Costa-Montenegro E, González-Castã FJ (2015) GTI: an unsupervised approach for sentiment analysis in twitter. In: Proceedings of 9th international workshop on semantic evaluation (SemEval 2015), pp 533–538
Fernández-Gavilanes M, Álvarez-López T, Juncal-Martínez J, Costa-Montenegro E, Javier González-Castaño F (2016) Unsupervised method for sentiment analysis in online texts. Expert Syst Appl 58:57–75
Gerber MS (2014) Predicting crime using Twitter and kernel density estimation. Decis Support Syst 61(1):115–125
Ghosh R, Ravi K, Ravi V (2017) A novel deep learning architecture for sentiment classification. In: 3rd International conference on recent advances in information technology|RAIT-2016|, pp 3–8
Giachanou A, Crestani F (2016) Like it or not: a survey of twitter sentiment analysis methods. ACM Comput Surv 49(2):28:3–28:40
Go A, Bhayani R, Huang L (2009) Twitter sentiment classification using distant supervision. CS224 N Proj Rep Stanf 1(12):1–6
Hafez G, Ismail R, Karam O (2017) Temporal sentiment analysis and time tags for opinions. In: The 8th IEEE international conference on intelligent computing and information systems (ICICIS 2017), pp 373–378
Hakak NM, Mohd M, Kirmani M, Mohd M (2017) Emotion analysis: a survey. In: International conference on computer, communications and electronics, COMPTELIX 2017, pp 397–402
Halin AA (2017) The importance of multimodality in sarcasm detection for sentiment analysis. In: IEEE 15th student conference on research and development (SCOReD), pp 56–60
Hao Y, Mu T, Hong R, Wang M, Liu X, Goulermas JY (2019) Cross-domain sentiment encoding through stochastic word embedding. IEEE Trans Knowl Data Eng 1–15
Haque TU, Saber NN, Shah FM (2018) Sentiment analysis on large scale amazon product reviews. In: IEEE international conference on innovative research and development (ICIRD), pp 1–6
Haselmayer M, Jenny M (2017) Sentiment analysis of political communication: combining a dictionary approach with crowdcoding. Qual Quant 51(6):2623–2646
Hassan A, Mahmood A (2017a) Efficient deep learning model for text classification based on recurrent and convolutional layers. In: 16th IEEE international conference on machine learning and applications (ICMLA), pp 1108–1113
Hassan A, Mahmood A (2017b) Deep learning approach for sentiment analysis of short texts. In: 3rd International conference on control, automation and robotics (ICCAR), pp 705–710
Hassan A, Mahmood A (2018) Convolutional recurrent deep learning model for sentence classification. IEEE Access 6:2169–3536
Hedge Y, Padma SK (2017) Sentiment analysis using random forest ensemble for mobile product reviews in Kannada. In: IEEE 7th international advance computing conference
Hemmatian F, Sohrabi M (2017) A survey on classification techniques for opinion mining and sentiment analysis. Artif Intell Rev 2017:1–51
Hogenboom A, Heerschop B, Frasincar F, Kaymak U, De Jong F (2014) Multi-lingual support for lexicon-based sentiment analysis guided by semantics. Decis Support Syst 62:43–53
Huang Q, Chen R, Zheng X, Dong Z (2017) Deep sentiment representation based on CNN and LSTM. In: Proceedings of 2017 international conference on green informatics, ICGI 2017, pp 30–33

Huang W, Rao G, Feng Z, Cong Q (2018) LSTM with sentence representations for document-level sentiment classification. Neurocomputing 308:49

Huang F, Zhang X, Zhao Z, Xu J, Li Z (2019) Image-text sentiment analysis via deep multimodal attentive fusion. Knowl Based Syst 167:26–37

Islam J, Zhang Y (2016) Visual sentiment analysis for social images using transfer learning approach. In: IEEE international conferences on big data and cloud computing (BDCloud), social computing and networking (SocialCom), sustainable computing and communications (SustainCom) (BDCloud-SocialCom-SustainCom), pp 124–130

Jaffali S, Jamoussi S, Ben Hamadou A (2014) Grouping like-minded users based on text and sentiment analysis. In: International conference on computational collective intelligence. Springer, Cham, pp 83–93

Jiang M, Wang J, Lan M, Wu Y (2014) An effective gated and attention-based neural network model for fine-grained financial target-dependent sentiment analysis. Int Conf Knowl Sci Eng Manag 214:42–54

Jin Y, Zhang H, Du D (2017) Improving deep belief networks via delta rule for sentiment classification. In: Proceedings of 2016 IEEE 28th international conference on tools with artificial intelligence, ICTAI 2016, pp 410–414

Jou B, Chen T, Pappas N, Redi M, Topkara M, Chang SF (2015) Visual affect around the world: a large-scale multilingual visual sentiment ontology. In: Proceedings of the 23rd ACM international conference on multimedia. ACM, pp 159–168

Kharde VA, Sonawane SS (2016) Sentiment analysis of twitter data: a survey of techniques. Int J Comput Appl 139(11):975–8887

Kim Y (2014) Convolutional neural networks for sentence classification. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), October 25–29, Doha, Qatar, pp 1746–1751

Kim J, Jang S, Park E, Choi S (2019) Text classification using capsules. Neurocomputing 118:247–261

Kiritchenko S, Zhu X, Mohammad S (2014) Sentiment analysis of short informal texts. J Artif Intell Res 50:723–762

Kraus M, Feuerriegel S (2019) Sentiment analysis based on rhetorical structure theory: learning deep neural networks from discourse trees. Expert Syst Appl 118:65–79

Krejzl P, Hourová B, Steinberger J (2017) Stance detection in online discussions. arXiv preprint arXiv:1701.00504

Kumari S, Babu CN (2017) Real time analysis of social media data to understand people emotions towards national parties. In: 8th International conference on computing, communication and networking technologies (ICCCNT), pp 1–6

Kušen E, Strembeck M (2017) Politics, sentiments, and misinformation: an analysis of the Twitter discussion on the 2016 Austrian presidential elections. Online Soc Netw Media 5:37–50

Lakkaraju H, Socher R, Manning CD (2014) Aspect specific sentiment analysis using hierarchical deep learning. In: NIPS workshop on deep learning and representation learning, pp 1–9

Lee G, Jeong J, Seo S, Kim CY, Kang P (2018) Sentiment classification with word localization based on weakly supervised learning with a convolutional neural network. Knowl Based Syst 152:70–82

Li H, Xu H (2019) Video-based sentiment analysis with hvnLBP-TOP feature and bi-LSTM. In: Association for the Advancement of Artificial Intelligence (AAAI)

Li C, Xu B, Wu G, He S, Tian G, Hao H (2014) Recursive deep learning for sentiment analysis over social data. In: Proceedings of 2014 IEEE/WIC/ACM international joint conference on web intelligence and intelligent agent technology–workshops, WI-IAT 2014, pp 180–185

Li Y, Pan Q, Yang T, Wang S, Tang J, Cambria E (2017a) Learning word representations for sentiment analysis. Cognit Comput 9(6):843–851

Li C, Guo X, Mei Q (2017b) Deep memory networks for attitude identification. In: Proceedings of the tenth ACM international conference on web search and data mining, WSDM 2017, Cambridge, United Kingdom, pp 671–680

Li B, Cheng Z, Xu Z, Ye W, Lukasiewicz T, Zhang S (2019) Long text analysis using sliced recurrent neural networks with breaking point information enrichment. In: Proceedings of the 2019 IEEE international conference on acoustics, speech and signal processing, ICASSP 2019, Brighton, UK, vol 124, pp 51–60

Liu B (2010) Sentiment analysis and subjectivity. In: Handbook of natural language processing, vol 1, pp 1–38

Liu Y, Bi J-W, Fan Z-P (2017) Ranking products through online reviews: a method based on sentiment analysis technique and intuitionistic fuzzy set theory. Inf Fusion 36:149–161

Lo S, Cambria E, Chiong R, Cornforth D (2017) Multilingual sentiment analysis: from formal to informal and scarce resource languages. Artif Intell Rev 48(4):499–527

Luo Z, Xu H, Chen F (2019) Audio sentiment analysis by heterogeneous signal features learned from utterance-based parallel neural network. In: Proceedings of the AAAI-19 workshop on affective content analysis, Honolulu, USA

Ma Y, Peng H, Khan T, Cambria E, Hussain A (2018) Sentic LSTM: a hybrid network for targeted aspect-based sentiment analysis. Cognit Comput 10:639–650

Ma X, Zeng J, Peng L, Fortino G, Zhang Y (2019) Modeling multi-aspects within one opinionated sentence simultaneously for aspect-level sentiment analysis. Futur Gener Comput Syst 93:304–311

Maas AL, Daly RE, Pham PT, Huang D, Ng AY, Potts C (2011) Learning word vectors for sentiment analysis. In: Proceedings of 49th annual meeting of the Association for Computational Linguistics: Human Language and Technology, pp 142–150

Manshu Y, Bing W (2019) Adding prior knowledge in hierarchical attention neural network for cross domain sentiment classification. IEEE Access 7:2169–3536

Marcheggiani D, Oscar T (2014) Hierarchical multi-label conditional random fields for aspect-oriented opinion mining. In: European conference on information retrieval. Springer, Cham, pp 273–285

Marelli M, Bentivogli L, Baroni M, Bernardi R, Menini S, Zamparelli R (2014) SemEval-2014 task 1: evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In: Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014), no 1, pp 1–8

Mataoui M, Hacine T, Tellache I, Bakhtouchi A, Zelmati O (2018) A new syntax-based aspect detection approach for sentiment analysis in Arabic reviews. In: 2nd international conference on natural language and speech processing (ICNLSP)

Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems, pp 3111–3119

Moghaddam S, Ester M (2010) Opinion digger: an unsupervised opinion miner from unstructured product reviews. In: Proceedings of the 19th ACM international conference on information and knowledge management, pp 1825–1828

Mohammad SM, Kiritchenko S, Zhu X (2013) NRC-Canada: building the state-of-the-art in sentiment analysis of tweets. In: Proceedings of the seventh international workshop on semantic evaluation, pp 321–327

Montejo-Ráez A, Díaz-Galiano MC, Martínez-Santiago F, Ureña-Løpez LA (2014) Crowd explicit sentiment analysis. Knowl Based Syst 69(1):134–139

Morency L-P, Mihalcea R, Doshi P (2011) Towards multimodal sentiment analysis: harvesting opinions from the web. In: Proceedings of 13th international conference on multimodal interfaces—ICMI'11, pp 169–176

Nakov P, Rosenthal S, Kozareva Z, Stoyanov V, Ritter A, Wilson T (2013) SemEval-2013 task 2: sentiment analysis in Twitter. In: Joint conference on lexical and computational semantics (SEM). Volume 2: Proceedings of the international workshop on semantic evaluation (SemEval 2013), vol 2, no SemEval, pp 312–320

Napitu F, Bijaksana MA, Trisetyarso A, Heryadi Y (2017) Twitter opinion mining predicts broadband internet's customer churn rate. In: IEEE international conference on cybernetics and computational intelligence (CyberneticsCom), pp 141–146

Narr S, Hülfenhaus M, Albayrak S (2012) Language-independent twitter sentiment analysis. In: Workshop on knowledge discovery, data mining and machine learning (KDML-2012), Dortmund, Germany

Nogueira C, Santos D, Gatti M (2014) Deep convolutional neural networks for sentiment analysis of short texts. In: Proceedings of 25th international conference on computational linguistics, pp 69–78

Nozza D, Fersini E, Messina E (2016) Deep learning and ensemble methods for domain adaptation. In: IEEE 28th international conference on tools with artificial intelligence deep, pp 184–189

Pang B, Lee L (2004) A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the 42nd annual meeting on Association for Computational Linguistics, p 271

Pang B, Lee L, Vaithyanathan S (2002) Thumbs up? Sentiment classification using machine learning techniques. In: Empirical methods in natural language processing (EMNLP), vol 10, pp 79–86

Peñalver-Martinez I et al (2014) Feature-based opinion mining through ontologies. Expert Syst Appl 41(13):5995–6008

Peng H, Ma Y, Li Y, Cambria E (2018) Learning multi-grained aspect target sequence for Chinese sentiment analysis. Knowl Based Syst 148:55–65

Pennington J, Socher R, Manning CD (2014) GloVe: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1532–1543

P'erez-Rosas V, Mihalcea R, Morency L (2013) Utterance-level multimodal sentiment analysis. In: Proceedings of the 51st annual meeting of the Association for Computational Linguistics, pp 973–982

Pontiki M, Galanis D, Pavlopoulos J, Papageorgiou H, Androutsopoulos I, Manandhar S (2014) SemEval-2014 task 4: aspect based sentiment analysis. In: Proceedings of the 8th international workshop on semantic evaluation, pp. 27–35

Pontiki M et al (2016) SemEval-2016 task 5: aspect based sentiment analysis. In: Proceedings of the 10th international workshop on semantic evaluation, pp 342–349
Poria S, Cambria E, Howard N, Bin Huang G, Hussain A (2016a) Fusing audio, visual and textual clues for sentiment analysis from multimodal content. Neurocomputing 174:50–59
Poria S, Chaturvedi I, Cambria E, Hussain A (2016b) Convolutional MKL based multimodal emotion recognition and sentiment analysis. In: Proceedings-IEEE 16th international conference on data mining, ICDM, pp 439–448
Poria S, Cambria E, Gelbukh A (2016c) Aspect extraction for opinion mining with a deep convolutional neural network. Knowl Based Syst 108:42–49
Poria S, Cambria E, Bajpai R, Hussain A (2017a) A review of affective computing: from unimodal analysis to multimodal fusion. Inf Fusion 37:98–125
Poria S, Cambria E, Hazarika D, Majumder N, Zadeh A, Morency L-P (2017b) Context-dependent sentiment analysis in user-generated videos. In: Proceedings of the 55th annual meeting of the Association for Computational Linguistics (volume 1: long papers), pp 873–883
Poria S, Cambria E, Hazarika D, Mazumder N, Zadeh A, Morency LP (2017c) Multi-level multiple attentions for contextual multimodal sentiment analysis. In: Proceedings of IEEE international conference on data mining, ICDM, pp 1033–1038
Poria S, Majumder N, Hazarika D, Cambria E, Gelbukh A, Hussain A (2018) Multimodal sentiment analysis: addressing key issues and setting up the baselines. IEEE Intell Syst 33(6):17–25
Radianti J, Hiltz SR, Labaka L (2016) An overview of public concerns during the recovery period after a major earthquake: Nepal twitter analysis. In: Proceedings of the 49th annual Hawaii international conference on system sciences, pp 136–145
Ragini JR, Anand PMR, Bhaskar V (2018) Big data analytics for disaster response and recovery through sentiment analysis. Int J Inf Manag 42(May):13–24
Rain C (2013) Sentiment analysis in Amazon reviews using probabilistic machine learning. Swarthmore College
Rana TA, Cheah Y-N (2016) Aspect extraction in sentiment analysis: comparative analysis and survey. Artif Intell Rev 46(4):459–483
Rana R et al (2016) Gated recurrent unit (GRU) for emotion classification from noisy speech. arXiv preprint arXiv:1612.07778
Rani S, Kumar P (2019) A journey of Indian languages over sentiment analysis: a systematic review. Artif Intell Rev 52(2):1415–1462
Rao T, Srivastava S (2012) Analyzing stock market movements using Twitter sentiment analysis. In: ASONAM'12 Proceedings of the 2012 international conference on advances in social networks analysis and mining (ASONAM 2012), pp 119–123
Ravi K, Ravi V (2015) A survey on opinion mining and sentiment analysis: tasks, approaches and applications, vol 89. Elsevier, Amsterdam
Ren Y, Zhang Y, Zhang M, Ji D (2016) Context-sensitive twitter sentiment classification using neural network. In: Proceedings of the 30th conference on artificial intelligence (AAAI 2016), pp 215–221
Rosenfeld R, Fornango R (2008) The impact of economic conditions on robbery and property crime: the role of consumer sentiment. Criminology 45(4):735–769
Roy K, Kohli D, Kumar R, Sahgal R, Yu W-B (2017) Sentiment analysis of Twitter data for demonetization in India: a text mining approach. Inf Syst 18(4):9–15
Ruangkanokmas P, Achalakul T, Akkarajitsakul K (2016) Deep belief networks with feature selection for sentiment classification. In: 7th International conference on intelligent systems, modelling and simulation (ISMS), pp 9–14
Sabour S, Frosst N, Hinton GE (2017) Dynamic routing between capsules. In: Advances in neural information processing systems, pp 3856–3866
Saif H, Fernandez M, He Y, Alani H (2013) Evaluation datasets for Twitter sentiment analysis A survey and a new dataset, the STS-Gold. In: Proceedings of 1st ESSEM work, Turin, Italy, vol 1096, pp 9–21
Sánchez-rada JF, Iglesias CA (2019) Social context in sentiment analysis: formal definition, overview of current trends and framework for comparison. Inf Fusion 52:344–356
Shah RR, Yu Y, Verma A, Tang S, Shaikh AD, Zimmermann R (2016) Leveraging multimodal information for event summarization and concept-level sentiment analysis. Knowl Based Syst 108:102–109
Shaikh T, Deshpande D (2016) Feature selection methods in sentiment analysis and sentiment classification of amazon product reviews. Int J Comput Trends Technol 36(4):225–230
Shi S, Zhao M, Guan J, Li Y, Huang H (2017) A hierarchical LSTM model with multiple features for sentiment analysis of sina weibo texts. In: International conference on Asian language processing (IALP), pp 379–382

Singh P, Dave A, Dar K (2017) Demonetization: sentiment and retweet analysis. In: International conference on inventive computing and informatics (ICICI 2017), pp 894–899
Singh P, Sawhney RS, Kahlon KS (2018) Sentiment analysis of demonetization of 500 & 1000 rupee banknotes by Indian government. ICT Express 4:124
Singhal P, Bhattacharyya P (2016) Sentiment analysis and deep learning: a survey. In: Center for Indian Language Technology, Indian Institute of Technology, Bombay
Singla Z, Randhawa S, Jain S (2017) Statistical and sentiment analysis of consumer product reviews. In: 8th International conference on computing, communication and networking technologies (ICCCNT), pp 1–6
Socher R, Perelygin A, Wu J (2013) Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of 2013 conference on empirical methods in natural language processing, pp 1631–1642
Soleymani M, Garcia D, Jou B, Schuller B, Chang SF, Pantic M (2017) A survey of multimodal sentiment analysis. Image Vis Comput 65:3–14
Song K, Yao T, Ling Q, Mei T (2018) Boosting image sentiment analysis with visual attention. Neurocomputing 312:218–228
Stojanovski D, Strezoski G, Madjarov G, Dimitrovski I (2015) Twitter sentiment analysis using deep convolutional neural network. In: Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval. ACM, pp 726–737
Stojanovski D, Strezoski G, Madjarov G, Dimitrovski I, Chorbev I (2018) Deep neural network architecture for sentiment analysis and emotion identification of Twitter messages. Multimed Tools Appl 77(24):32213–32242
Sun X, Li C, Ren F (2016) Sentiment analysis for Chinese microblog based on deep neural networks with convolutional extension features. Neurocomputing 210:227–236
Tai KS, Socher R, Manning CD (2015) Improved semantic representations from tree-structured long short-term memory networks. arXiv preprint arXiv:1503.00075
Tang D, Qin B, Liu T (2015a) Document modeling with gated recurrent neural network for sentiment classification. In: Proceedings of the 2015 conference on empirical methods in natural language processing
Tang D, Qin B, Liu T (2015b) Learning Semantic representations of users and products for document level sentiment classification. In: Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing, vol 1, pp 1014–1023
Tay Y, Tuan LA, Hui SC (2017) Dyadic memory networks for aspect-based sentiment analysis. In: Proceedings of 2017 ACM conference on information and knowledge management—CIKM'17, pp 107–116
Trofimova TP, Pushin AN, Lys YI, Fedoseev VM (2016) Robust visual-textual sentiment analysis: when attention meets tree-structured recursive neural networks. In: Proceedings of the 2016 ACM on multimedia conference, pp 1008–1017
Twitter Statistics (2019). https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/
Uysal AK, Murphey YL (2017) Sentiment classification: feature selection based approaches versus deep learning. In: IEEE international conference on computer and information technology (CIT), pp 23–30
van Hee C, Lefever E, Hoste V (2018) Exploring the fine-grained analysis and automatic detection of irony on Twitter. Lang Resour Eval 1–25
Vateekul P, Koomsubha T (2016) A study of sentiment analysis using deep learning techniques on Thai Twitter data. In: 13th International joint conference on computer science and software engineering (JCSSE), pp 1–6
Verma S, Saini M, Sharan A (2018) Deep sequential model for review rating prediction. In: 10th international conference on contemporary computing, IC3 2017, vol 2018, pp 1–6
Wang H, Can D, Kazemzadeh A, Bar F, Narayanan S (2012a) A system for real-time twitter sentiment analysis of 2012 U.S. presidential election cycle. In: Proceedings of the 50th annual meeting of the Association for Computational Linguistics, pp 115–120
Wang X, Gerber MS, Brown DE (2012b) Automatic crime prediction using events extracted from twitter posts
Wang Y, Huang M, Zhu X, Zhao L (2016a) Attention-based LSTM for aspect-level sentiment classification. In: Proceedings of the 2016 conference on empirical methods in natural language processing, pp 606–615
Wang H, Meghawat A, Morency L, Xing EP (2016b) Select-additive learning: improving generalization in multimodal sentiment analysis. arXiv preprint arXiv:1609.05244
Wang J, Fu J, Xu Y, Mei T (2016c) Beyond object recognition: visual sentiment analysis with deep coupled adjective and noun neural networks. In: Proceedings of the twenty-fifth international joint conference on artificial intelligence (IJCAI-16), pp 3484–3490
Wang X, Li Y, Xu P (2018a) A hybrid BLSTM-C neural network proposed for chinese text classification. In: IEEE sixth international conference on advanced cloud and big data (CBD), pp 311–315
Wang Y, Sun A, Han J, Liu Y, Zhu X (2018b) Sentiment analysis by capsules. In: Proceedings of the 2018 world wide web conference, pp 1165–1174

Wang J, Peng B, Zhang X (2018c) Using a stacked residual LSTM model for sentiment intensity prediction. Neurocomputing 322:93–101
Wang Y, Sun A, Huang M, Zhu X (2019) Aspect-level sentiment analysis using AS-capsules. In: The world wide web conference. ACM, pp 2033–2044
Whitehead M, Yaeger L (2009) Building a general purpose cross-domain sentiment mining model. In: WRI world congress on computer science and information engineering, CSIE, vol 4, pp 472–476
Wu D, Chi M (2017) Long short-term memory with quadratic connections in recursive neural networks for representing compositional semantics. IEEE Access 5:16077
Wu D, Cui Y (2018) Disaster early warning and damage assessment analysis using social media data and geo-location information. Decis Support Syst 111:48
Xiong S, Wang K, Ji D, Wang B (2018a) A short text sentiment-topic model for product reviews. Neurocomputing 297:94–102
Xiong S, Lv H, Zhao W, Ji D (2018b) Towards Twitter sentiment classification by multi-level sentiment-enriched word embeddings. Neurocomputing 275:2459–2466
Xu F, Kešelj V (2014) Collective sentiment mining of microblogs in 24-hour stock price movement prediction. In: 16th IEEE conference on business informatics, CBI 2014, vol 2, pp 60–67
Xu K, Liao SS, Li J, Song Y (2011) Mining comparative opinions from customer reviews for competitive Intelligence. Decis Support Syst 50(4):743–754
Xu J, Tao Y, Lin H, Zhu R, Yan Y (2017) Exploring controversy via sentiment divergences of aspects in reviews. In: IEEE pacific visualization symposium (PacificVis), pp 240–249
Yanagimoto H, Shimada M, Yoshimura A (2013) Document similarity estimation for sentiment analysis using neural network. In: IEEE/ACIS 12th international conference on computer and information science (ICIS). IEEE, pp 105–110
Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E (2016) Hierarchical attention networks for document classification. In: Conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies, pp 1480–1489
Yang M, Qu Q, Chen X, Guo C, Shen Y, Lei K (2018) Feature-enhanced attention network for target-dependent sentiment classification. Neurocomputing 307:91–97
Yang C, Zhang H, Jiang B, Li K (2019a) Aspect-based sentiment analysis with alternating coattention networks. Inf Process Manag 56(3):463–478
Yang M, Zhao W, Chen L, Qu Q, Zhao Z, Shen Y (2019b) Investigating the transferring capability of capsule networks for text classification. Neural Netw 118:247–261
Yelp Dataset (2014)
Yoo SY, Song JI, Jeong OR (2018) Social media contents based sentiment analysis and prediction system. Expert Syst Appl 105:102–111
You Q, Luo J, Jin H, Yang J (2015) Robust image sentiment analysis using progressively trained and domain transferred deep networks. In: Proceedings of the twenty-ninth AAAI conference on artificial intelligence, pp 381–388
You Q, Luo J, Jin H, Yang J (2016) Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia. In: Proceedings of the ninth ACM international conference on web search and data mining, pp 13–22
Yu H, Gui L, Madaio M, Ogan A, Cassell J (2017) Temporally selective attention model for social and affective state recognition in multimedia content. In: Proceedings of the 2017 ACM on multimedia conference. ACM, pp 1743–1751
Yu L, Wang J, Lai KR, Zhang X (2018) Refining word embeddings using intensity scores for sentiment analysis. IEEE/ACM Trans Audio Speech Lang Process 26(3):671–681
Yu J, Jiang J, Xia R (2019) Global inference for aspect and opinion terms co-extraction based on multi-task neural networks. IEEE/ACM Trans Audio Speech Lang Process 27(1):168–177
Yuan M, Tang H, Li H (2014) Real-time keypoint recognition using restricted boltzmann machine. IEEE Trans Neural Netw Learn Syst 25(11):2119–2126
Yuan Z, Wu S, Wu F, Liu J, Huang Y (2018) Domain attention model for multi-domain sentiment classification. Knowl Based Syst 155:1–10
Zadeh A, Zellers R, Pincus E, Morency L (2016) MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *IEEE Intell Syst*
Zhang J, Chow C (2019) MOCA: multi-objective, collaborative, and attentive sentiment analysis. IEEE Access 7:10927–10936
Zhang Y, Wallace B (2015) A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. arXiv preprint arXiv:1510.03820
Zhang Z, Ye Q, Zhang Z, Li Y (2011) Sentiment classification of internet restaurant reviews written in Cantonese. Expert Syst Appl 38(6):7674–7682

Zhang Z, Zou Y, Gan C (2017) Textual sentiment analysis via three different attention convolutional neural networks and cross-modality consistent regression. Neurocomputing 275:1407

Zhang Y et al (2018a) A quantum-inspired multimodal sentiment analysis framework. Theor Comput Sci 752:21

Zhang Z, Wang L, Zou Y, Gan C (2018b) The optimally designed dynamic memory networks for targeted sentiment classification. Neurocomputing 309:36

Zhang B, Xu X, Yang M, Chen X, Ye AY (2018c) Cross-domain sentiment classification by capsule network with semantic rules. IEEE Access 6:58284–58294

Zhao L, Huang M, Chen H, Cheng J, Zhu X (2014) Clustering aspect-related phrases by leveraging sentiment distribution consistency. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1614–1623

Zhao W et al (2017) Weakly-supervised deep embedding for product review sentiment analysis. IEEE Trans Knowl Data Eng 4347:1–12

Zhao W, Peng H, Eger S, Cambria E, Yang M (2019) Towards scalable and reliable capsule networks for challenging NLP applications. In: Proceedings of the 57th annual meeting of the Association for Computational Linguistics, pp 1549–1559

Zhou K, Zeng J, Liu Y, Zou F (2018) Deep sentiment hashing for text retrieval in social CIoT. Futur Gener Comput Syst 86:362

Zvarevashe K, Olugbara OO (2018) A framework for sentiment analysis with opinion mining of hotel reviews. In: Conference on information communications technology and society (ICTAS), pp 1–4

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Short Term GA-NFIS based Hybrid Method for Prediction of Wind Speed & Power in Sustainable Power Generation

Vijay Kumar, Yash Pal, Madan Mohan Tripathi

Abstract*: Additional power requirement and integration of non conventional energy sources for power generation into smart grid, forced the world power generating company to divert their attention towards power generation from renewable sources along with conventional sources. Today power demand in different sectors are increasing with continuous increasing population, so it is very challenging task for power generating company to maintain balance between supply and demand. The countries that consume more energy per capita are supposed to have better social- economical status and living standard. Currently power generation from the conventional sources faces many challenges such as environment pollution, continuous availability, storage and security; to mitigate the above problems power generation from renewable sources may be better alternative. So, many countries started power generation from renewable sources such as wind power and solar power by expanding major portion of their energy fund into development of the renewable setup for power generation. In India currently wind power contribution in renewable power sources is very high so, wind may be a better solution for power generation. Power generation from wind puts many barriers in term of intermittent nature, frequency and availability at all places with certain speed that is able to power generation from wind mills. Although above challenge cannot eliminate completely, but it can be minimize with the help of correct prediction technique / method that have accuracy up to certain level which will be acceptable for power generation. In this work GA- NFIS is used for forecasting of wind power and data is collected from Indian wind power sector mills. Results of proposed method are compared with some available soft computing methods such as NARX, SVM-NARX.

Index Terms*: Prediction, Wind power, Genetic Algorithm, Neuro- Fuzzy Inference System, Speed prediction, Absolute Percentage Error (APE).

I. INTRODUCTION

With rapid growing population along with industrialization in the world currently power demand and its proper supply from different sources has become serious issue for the whole country of the world. To fulfill the current requirement of energy we cannot depend only on conventional method of power generation or sources due to its limited storage, pollution. So, there will be requirement of balance between demand and supply while insuring the minimum possible pollution. By increasing the contribution of renewable share in power generation above problems can be reduce up to some extent. For this purpose wind power may be one of the good alternative solutions [1]. The main problem of power generation from wind is its unpredictable nature with varying speed and it can be minimize up to certain acceptable level with suitable prediction of wind speed with good accuracy that will enhance power generation economically. Due to highly dependency of wind power on speed, very small error in speed prediction will reflect considerable error in power calculation. Keeping the above point there will be requirement of such types of technique / method that will predict speed with minimum error [2]. Now days many different methods/ techniques are available in the field of forecasting such as ARMA / ARIMA model, Persistence method, Physical, Statistical method, NWP, AI based methods. Depending on situation and requirement of time interval different methods are utilized [4] [5]. Paper is distributed into 7 section such as section 1 give the brief introduction, section 2 discuss about genetic algorithm (GA), section 3 provide basic information about Neuro-Fuzzy Inference System (NFIS), section 4 describes the forecasting procedure of suggested method by GA & NFIS, section 5 provide the knowledge of basic error parameters related to prediction, section 6 shows results, comparison with others method and last section tells the conclusion. The result comparison indicates that proposed method provide better results compare to other.

II. ARTIFICIAL INTELLIGENCE BASED METHOD (ANN)

Before the regulated electricity markets Neural network was primarily applied on load & price forecasting, but currently it is also apply in the field of wind parameters prediction like speed and power [6]. ANN has capability to save the computing time by minimizing numbers of variables during training process [7]. ANN provides better result for time series data in different time slots with respect of other conventional methods. In modeling time series data generally we use nonlinear autoregressive exogenous model (NARX), in which present exogenous based input value reflect a relationship to the previous value. NARX model can be mathematically modeled through equation 1.

$$R_m = P\,(x_{t-1,}\, x_{t-2,}\, x_{t-3\ldots}\, m_{t,}\, m_{t-1},\, m_{t-2,}\, m_{t-3\ldots}) + E_t \qquad (1)$$

Here x_i represent variable of interest, *m* determination variable & E_t error. Basic Multilayer (3- layer) ANN is shown in figure 1 this network comprises of different layers such as inputs, hidden and corresponding to output. Modeling parameters of the network can be expressed by expression.

$$(R, t) = (R^{(1)}, t^{(1)}, R^{(2)}, t^{(2)}) \qquad (2)$$

Revised Manuscript Received on December 22, 2018.

Mr. V Kumar, National Institute of Technology, Kurukshetra, (Haryana), India.

Dr. Yash Pal, Professor In The EED, National Institute of Technology, Kurukshetra (Haryana), India.

Dr. Madan Mohan Tripathi, Professor in the EDD, Delhi Technological University, New Delhi

$R_{mn}^{(l)}$ reflect corresponding associated weight of units existing in layer l, with the corresponding units of m+1 layer, $t_m^{(l)}$ represent associated bias of unit m in m+1 layer. The ANN structure takes the previous data and current trends to predict the future data [14].

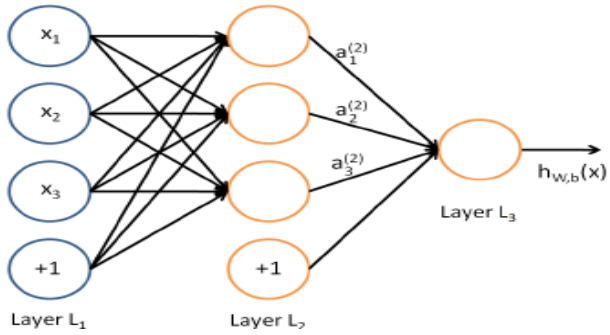


Fig.1: Multi-layer ANN

III. GENETIC ALGORITHMS (GA)

In the field of AI applications Genetic algorithm may be one of the most efficient and better computing technique/method in case of forecasting [8]. It works on principle of random searches technique with the help certain set of existing alternatives in attempt to find out best possible alternative based on the provided criterion. The different set of criteria is represents via an objective function that also known as fitness function. Working/ searching procedure for optimization condition of GA is represented in figure 2.

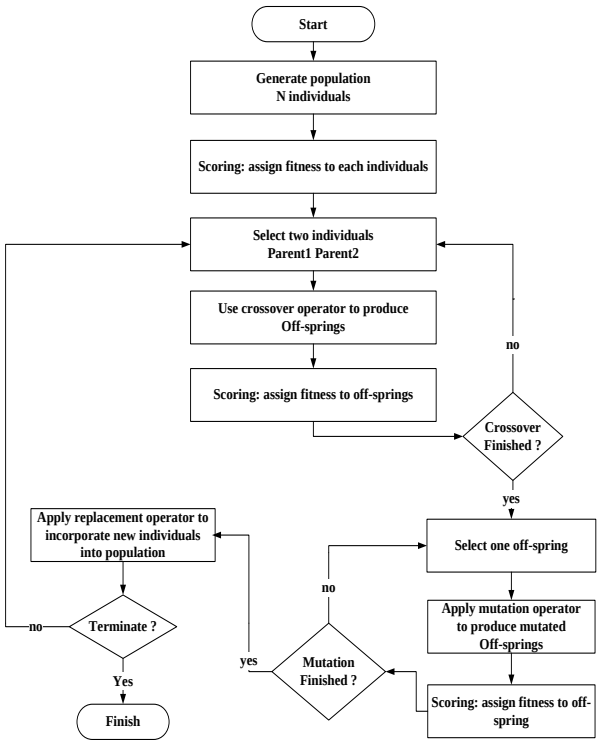


Fig.2. Working Procedure of Genetic algorithm

Suppose objective function is represented by G (n), which is to be optimized, suppose k_i , p_i is a constant and input variables respectively. Main aim of objective function is such that the value of equation represented by equation 3 will provide optimized value.

$$k\ (p_1, p_2 \ldots\ldots\ldots\ldots p_9) \qquad (3)$$

Where k_i (p) ≥ 0, for every i=1, 2, 3…n. Expression represented via k (p_1, p_2…….p_9) is variables. During processing GA sustained a population of G individuals such that Pop (k) = (p_1 (k),...,p_G (k)) for each calculation iteration there will k (generation variables) even single individual has capability to represents a good result for related objective function. In consecutive step 2 chromosomes may be taken as parent & next populations will be generating by crossover and mutation of above two chromosomes. With generation of new population, new fitness function again calculates for each set of chromosome. Out of this best fitness function and objective function are selected as parent chromosome for next calculation / generation. In this process with help of crossover operator, 2 new off springs are generated. Crossover operator generally applied on those particular members that have high probability of fitness. This process is repetitive and will repeat till required / termination criteria meet.

IV. NEURO – FUZZY INFERENCE SYSTEMS (NFIS)

NFIS provide better solution for such types of problems that is related to approximation of functions & may be regarded ANN in case of data driven process [9-11]. During nonlinear forecasting it takes the previous data to predict future data via including the self-learning ANN characteristics as well as NFIS function approximation procedure [12-13]. At the stage of final output estimation weighted data of each participant variables are consider. Figure 3 shows the basic structure of NFIS.

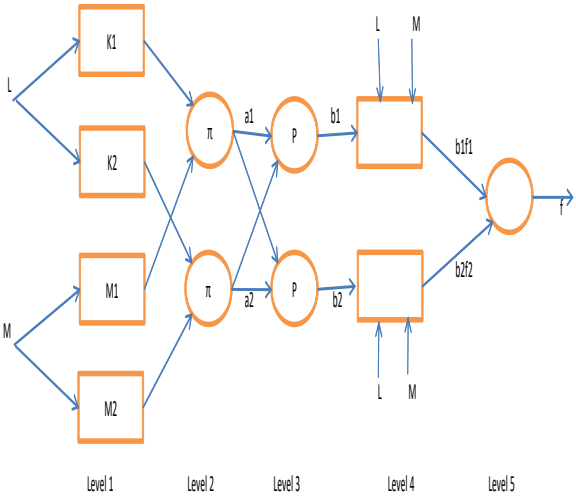


Fig.3. Basic structure of NFIS

NFIS network structure includes 5 corresponding internal layers in which each separate layer composed many numbers of nodes which is represented via node function. Functions as well as description details of each layers are described as below.

Layer 1: In this layer every node *n* is belong to other node via function that is given by equation 4.

$$P_n^1 = \mu M_n(R) \quad (4)$$

Where R is input related to node *n*, M_n label corresponding node n. Membership functions related to input variables are represented by equation 5.

$$\mu m_j(R) = \frac{1}{1+\left|\frac{R-r_j}{m_j}\right|^{2b_j}} \quad (5)$$

here R is input and $\{k_n, b_n, r_n\}$ is corresponding related parameter of node.

Layer 2: Individual node is fixed type node, it is responsible for rule based estimation of firing strength a_n. Output related to each node may be product of each input signals that is denoted by equation. (6).

$$P_n^2 = a_j = \mu M_n(R) = \mu M_n(m) \qquad n = 1, 2, 3 \quad (6)$$

Layer 3: Like layer 2, every node is fixed. For the calculation of total firing strengths of combined nodes each n^{th} node check the ratio of n^{th} node . Final output strength some times called as normalized strengths that can be represented by equation (7).

$$p_n^3 = \overline{b_n} = \frac{a_n}{a_1 + a_2}, \text{ Where n = 1, 2, 3,…} \quad (7)$$

V. ERROR PARAMETERS ESTIMATION

Accuracy of predicted technique / model is examined by estimation of different error parameters related to forecasting such as RSME, Mean bias error and Root Mean Square Error. Estimated parameters can be represented through equation 8, 9 and 10.

(i) RSME can calculated as

$$RSME = \sqrt{\frac{1}{K}\sum_{i=1}^{K}(P_{ki} - P_{ci})^2} \quad (8)$$

(ii) MBE can calculate

$$MBE = \frac{1}{K}\sum_{i=1}^{K}(P_{ki} - P_{ci}) \quad (9)$$

(iii) MAE can calculate as

$$MAE = \frac{1}{B}\sum_{i=1}^{n}\left|(P_{mi} - P_{ci})\right| \quad (10)$$

Here P_{ki}, P_{ci} is the actual power & forecasted power related to i^{th}, n represent the number of times that observation taken.

VI. PROPOSED TECHNIQUE

Currently proposed technique utilizes a hybrid method based on GA, and NFIS. First stage available input wind data's are normalized via in data processing, after that this normalized data is used for training and validation of proposed GA-NFIS model. Predicted values of prepared data sets for particular interval depending on requirement are forecasted by proposed GA- NFIS method. During forecasting process genetic algorithm (GA) is used to enhance final performance associated NFIS structure. After that final predicted wind data is used for the error estimation that will reflect the correctness of proposed method. Proposed working model of GA-NFIS is represented by figure 4.

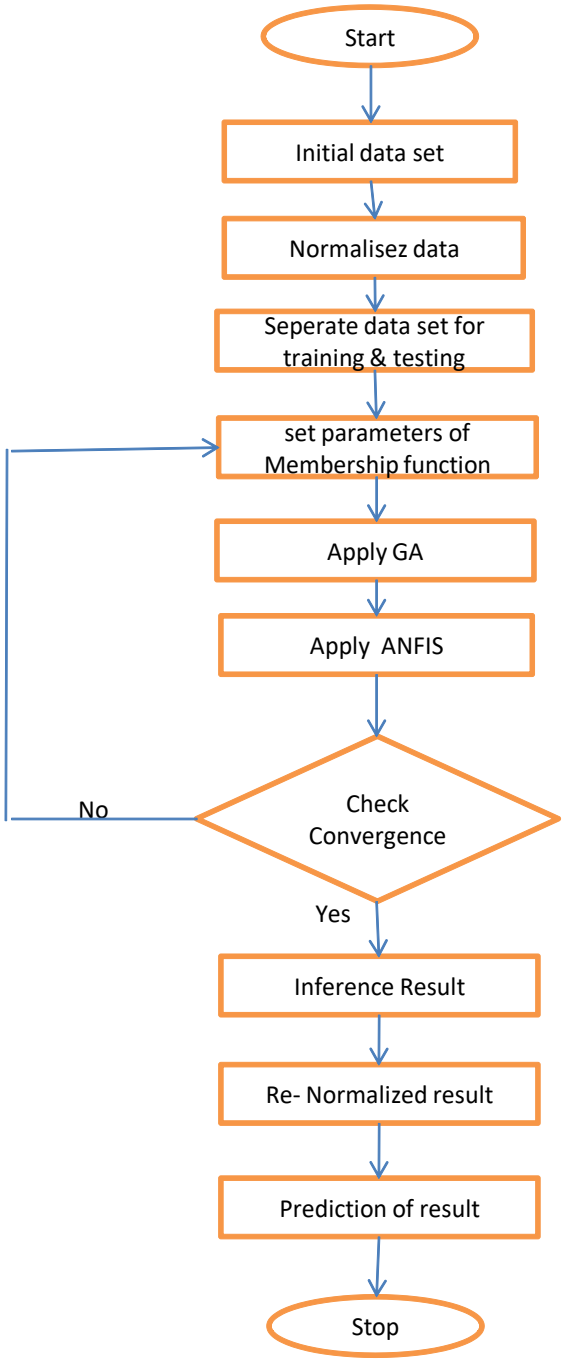


Fig.4. Proposed GA-NFIS model

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

VII. FORECASTED RESULTS

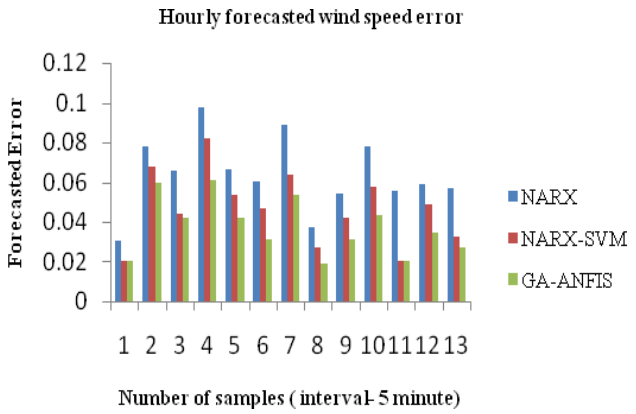


Fig. (5a) Forecasted error in speed

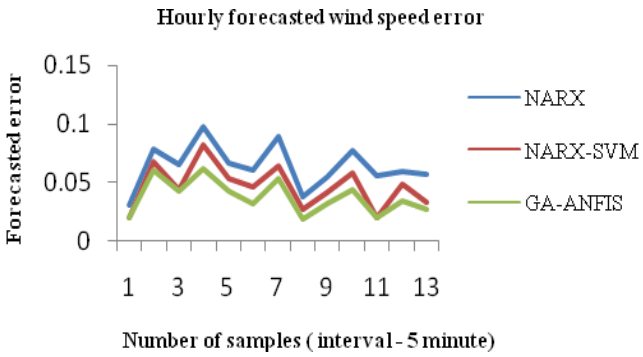


Fig. (5b) Hourly Forecasted error in speed

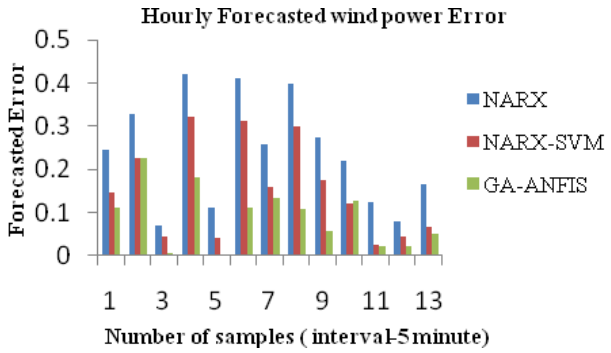


Fig. (6a) Hourly Forecasted error in power

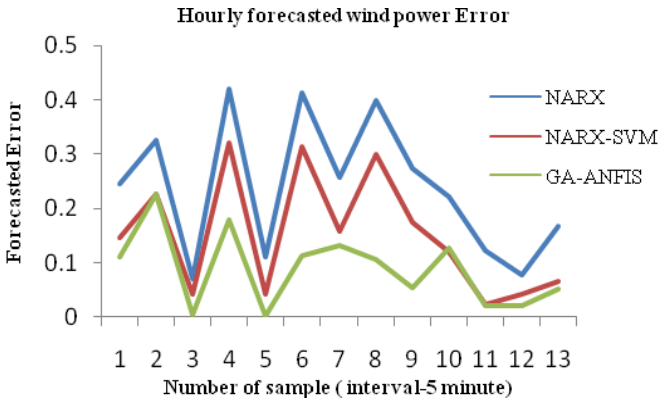


Fig. (6b) Forecasted error in power.

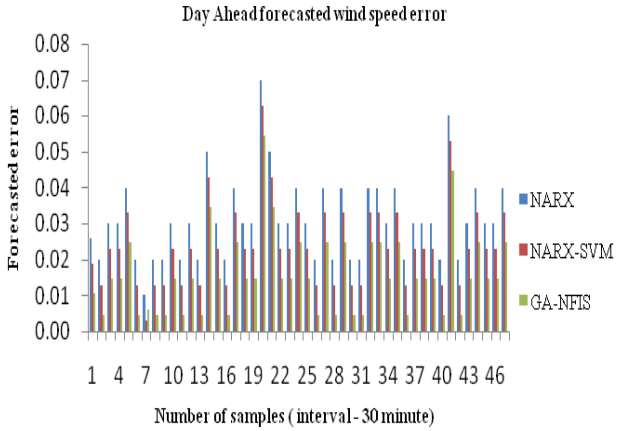


Fig. 7 Forecasted error in speed

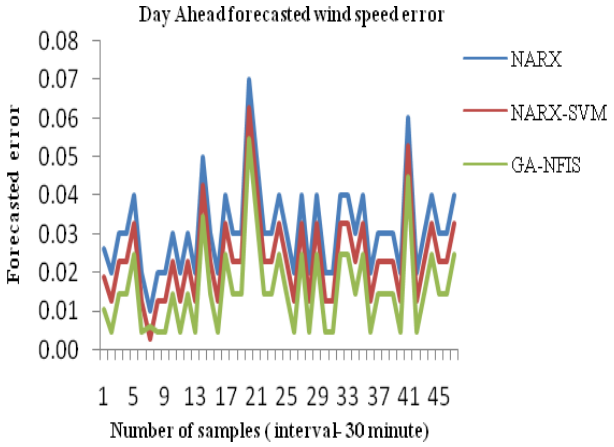


Fig. 8 Error prediction in speed

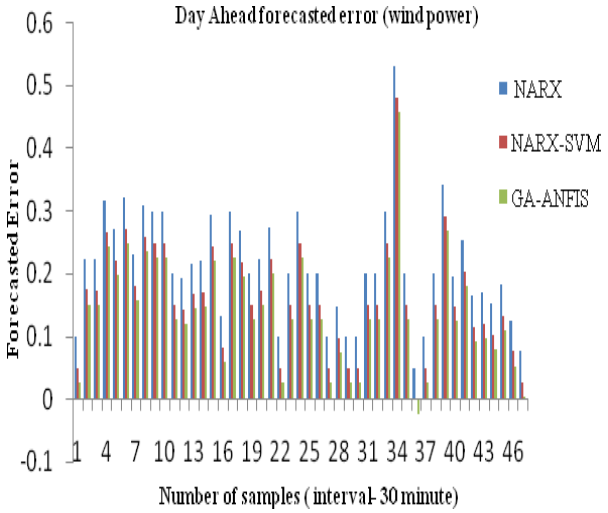


Fig. 9 Error prediction in power

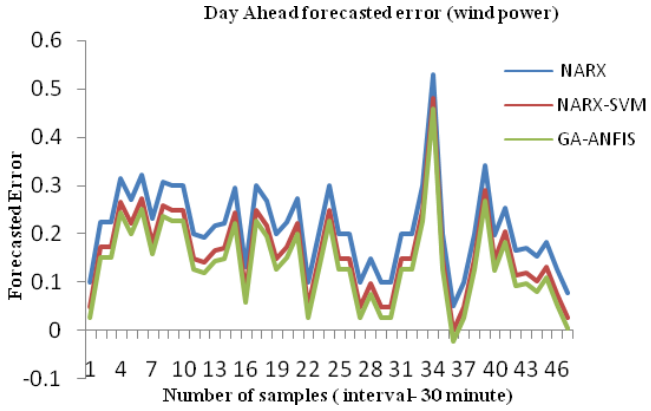


Fig. 10 Day Ahead error in wind power prediction

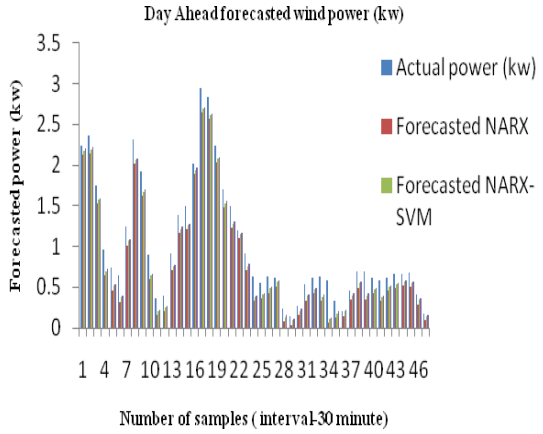


Fig. 11 Day Ahead predicted wind power

Table 1 & 2 shows the corresponding MAPE for wind speed is & power for NARX, NARX-SVM & GA-NFIS.

Table 1. PERCENTAGE MAPE FOR SPEED

Prediction Technique	Hourly MAPE	Day Ahead	Weekly MAPE	Monthly MAPE
NARX	7.28	8.57	9.24	9.55
NARX-SVM	5.79	6.22	7.33	8.18
GA-ANFIS	4.17	5.46	5.68	6.19

Table 2 PERCENTAGE MAPE FOR POWER

Prediction Technique	Hourly MAPE	Day Ahead	Weekly MAPE	Monthly MAPE
NARX	8.48	9.69	9.78	9.88
NARX-SVM	6.69	7.22	7.37	7.79
GA-ANFIS	4.79	5.85	6.34	6.86

VIII. CONCLUSION

With help of accurate and reliable prediction method that will provide minimum error between actual and predictedvalue we can increase the reliability and sustainability of wind power generation and its integration into smart grid that will help in power deficiency mitigation. The current proposed GA-NFIS based hybrid method is apply on the wind data obtained from Indian wind energy form for short term prediction and analysis of results shows good improvement in term of MAPE parameter as compare to NARX, NARX-SVM and few other methods. Comparative analysis of 3 different methods i.e NARX, NARX-SVM and GA-NFIS are shown in table 1 & 2. From the table 1, table 2 and figure 5 to 11 it is clear that proposed method i.e GA-NFIS give better result in term of error parameter MAPE. For hourly, day ahead, weekly and monthly wind speed prediction GA-NFIS provide MAPE 4.16, 5.46, 5.68 and 6.19 respectively, while for wind power prediction it is 4.79, 5.85, 6.34 and 6.86 respectively which better than other two methods for same prediction time interval.

REFERENCES

1. Georgilakis S. Pavlos, "The Technical challenges related with integration of wind power into power systems," Renewable &. Sustainable Energy Rev. Vol. 12, No.3, pp. 851– 862, 2008.
2. D K Soon, "Uncertainty analysis related assessment of wind energy potential," Applied Energy (2010), vol. 87, no 3, pp. 855–864.
3. M. Lei, J. Chuanwen. "A review on the forecasting of wind speed and generated power." Renewable and Sustainable Energy Review vol.13, issue 4, pp. 915-920, 2009.
4. N chen, Z Qian "Wind Forecasting Using Gaussian Processes (GP) and Numerical Weather Prediction (NWP)," IEEE Transaction on Power System Vol.29, no..2, pp. 655-666, 2014.
5. Marvuglia A, Foley A M, Leahy P G "Current methods and advances in forecasting of wind power generation". Renewable Energy Vol. 37, Issue 1, Jan. 2012, pp. 1-8.
6. Catalão J. P. D. S. P, Mendes V.MF. "Short-Term Wind Power Forecasting in Portugal by Neural Network Approach and wavelet transform" Renewable Energy vol.36, issue. 4, pp. 1245-1251, 2011.
7. C .D Dumitru and A. Gligor, "Daily average wind energy forecasting using artificial neural networks." Procedia Engineering 181(2017), pp. 829-836.
8. Jang. S, Sun C T. & Mizutani, "A Computational Approach to Learning and Machine Intelligence". IEEE Transactions on automatic control, volume 42, no. 10, October 1997, pp. 1482-1484.
9. Jang. S & Sun C. T, " Neuro-fuzzy modeling and control," Proceeding of .IEEE, Volume 83, Issue 3, March 1995, pp. 378–406.
10. H K Yadav, Y Pal & M M Tripathi, "A novel GA-NFIS hybrid model for short term solar power forecasting in Indian electricity market". Journal of Information and optimization science vol.22, issue 2, pp 377-395, 2019.
11. Yun Z , Quan Z, Caixin S & S. Yang, "RBF Neural Network and ANFIS-based short-term load forecasting approach in real-time price environment," IEEE Transaction on Power Systems, Volume. 23, Issue 3, Aug. 2008, pp. 853–858.
12. Shoorehdeli M.A, Teshnehlab M, &. Sedigh A. K, "Identification using ANFIS with intelligent hybrid stable learning algorithm approaches methods," Applied Computing, Vol. 9, Issue 2, March 2009, pp. 833–850.
13. G. A. Vijayalakshmi and S. Rajsekaran, "Neural Networks, Fuzzy Logic, and Genetic Algorithms synthesis and application," PHI Private limited, 14th Edition, January 2010.
14. V. Kumar, Y. Pal, MM Tripathi. "A hybrid SVM-NARX prediction method for Indian wind power sector" Journal of Statistics & Management system (JSMS), Vol. 22, Issue 2, 2019.

AUTHORS PROFILE

Mr. V Kumar is pursuing PhD from school of renewable energy & efficiency National Institute of Technology Kurukshetra . He completed under graduate (B.tech) , M.Tech from IERT Allahabad and National Institute of Technology Kurukshetra. Currently he is working in the field of soft computing technique, Renewable energy and Forecasting techniques related to wind and solar.

Dr. Y Pal is presently working as a Professor in the EED, National Institute of Technology Kurukshetra. He is currently working in the field of Power Quality, FACTS devices, Renewable Energy & Custom Power Devices.

Dr. Madan Mohan Tripathi is presently working as a Professor in the EDD, Delhi Technological University New Delhi. He is presently working in the field of Restructuring of power system, Soft computing technique, load & price prediction in deregulated electricity market, Security issues related to Smart Grid.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Skeleton based Activity Recognition by Fusing Part-wise Spatio-temporal and Attention Driven Residues

Chhavi Dhiman[1], Dinesh Kumar Vishwakarma[2], Paras Aggarwal[3]
Biometric Research Laboratory, Department of Information Technology,
Delhi Technological University, Delhi-110042
Email: chhavi1990delhi@gmail.com[1], dinesh@dtu.ac.in[2] , paraspro2020@gmail.com

Abstract:
There exist a wide range of intra-class variations of the same actions and inter-class similarity among the actions, at the same time, which makes the action recognition in videos very challenging. In this paper, we present a novel skeleton-based part-wise Spatio-temporal CNN – RIAC Network-based 3D human action recognition framework to visualise the action dynamics in part wise manner and utilise each part for action recognition by applying weighted late fusion mechanism. Part-wise skeleton-based motion dynamics helps to highlight local features of the skeleton which is performed by partitioning the complete skeleton in five parts- Head to Spine (HS), Left Leg (LL), Right Leg (RL), Left Hand (LH), Right Hand (RH). The RIAFNet architecture is greatly inspired by the InceptionV4 architecture which unified the ResNet and Inception based Spatio-temporal feature representation concept and achieving the highest top-1 accuracy till date. To extract and learn salient features for action recognition, attention driven residues are used which enhance the performance of residual components for effective 3D skeleton-based Spatio-temporal action representation. The robustness of the proposed framework is evaluated by performing extensive experiments on three challenging datasets such as UT Kinect Action 3D, Florence 3D action Dataset, and MSR Daily Action3D datasets, which consistently demonstrate the superiority of our method.

Keywords: Human Action Recognition, Skeleton, Attention, Residues, Inception, Spatio-temporal action representation.

1. Introduction

Human action recognition in videos is an essential field of computer vision and has an extensive range of applications i.e. Human-robot vision, human-computer interaction, intelligent video surveillance, game control, elderly health care, sports etc. [1] [2] [3]. Traditional researches in this field initially referred to RGB video-based action recognition. The information recorded in these videos is one projection of the entire scene, along with the motion content of the object

in action, on a particular plane. However, the motion characteristics of the human body are well represented in the 3D space. Considering this fact human bodies are, now, modelled as a set of 20 or 25 skeleton joints in 3D coordinate system and rigid bones. Position variation of the skeleton joints in 3D space are responsible for human actions. Currently, the world skeleton joint coordinates can be extracted using the cost-effective depth sensor combining the real-time skeleton estimation algorithms [4], [5]. It has resulted in significant amount of work, coming up by exploring effective approaches for skeleton-based action recognition. Human skeleton-based action recognition approaches generally exploit temporal dynamics of the action [6], by developing explicit temporal dynamics model such as Fourier Temporal Pyramids (FTPs) [7] [8] and probabilistic graphical model- HMMs [9] . FTP based methods utilize a limited amount of contextual information constrained by the size of the time windows. Hence, failed to capture the temporal sequences of actions, globally. Alignment of temporal sequences in HMM and obtaining a corresponding emission distributions is tough. Recently, several Recurrent Neural Networks (RNNs) models [10] and Long-Short Term Memory (LSTMs) models [11] are proposed to learn the temporal dynamics of action features. A large amount of information lies in the spatial distribution of human poses [12] [13]. Well-designed spatial-temporal encodings of skeleton sequences are more effective than only temporal information of skeleton-based action representation. Therefore, the methods [14] [15] [16] are proposed which process both spatial and temporal information of the action together, leading to more effective recognition systems. Pham et al. [14] [15] transformed the body joint coordinates into 3D arrays which can capture the spatial-temporal evolutions of 3D motions from skeleton sequences and processed as images using Deep Residual Neural Networks (ResNets) to learn and recognize human action from skeleton data provided by Kinect sensor.

In this work, we aim to take full advantages of spatio-temporal CNN to build an end-to-end learning framework for HAR by utilising both spatial and temporal information of action using 3D skeleton based representations of the actions. The concept of InceptionV4 [17] architecture has evolved from InceptionV1/ GoogleNet [18] and later fused with the concept of ResNet [19]. It greatly enhanced the recognition performance of the InceptionV4 [17] architecture from other versions of Inception network with the minimum top-1 error. However, in order to achieve superior performance than other state-of-the-art, use of InceptionV4 [17] network significantly demands a more significant number of parameters and operations. It greatly inspired us, to utilise the concept of inception blocks and residues, in designing the proposed RIAC-Net architecture, for feature extraction. The proposed work defines a novel action descriptor by combining

attention based residues with Spatial-Temporal based Convolution Features (STCF). The main contribution of the proposed work are four folds: -

- A novel part wise Spatio-temporal CNN based action descriptor –RIAC-Net is designed, using attention driven residual concept. The complete skeleton is partitioned into five significant parts: a) head to spinal (HS), b) left leg (LL), c) right leg (RL), d) left hand (LH), and e) right hand (RH).
- To learn the temporal dynamics of the complete action, the defined part-wise action descriptors, are learnt using two consecutive layers of Long Short Term Memory (LSTM) units. And the final prediction for the test action is taken by using a weighted fusion of five predictions computed from each part of the skeleton.
- To exhibit the effect of the engineering part-wise skeleton-based RIAC-Net action description over complete skeleton-based action description, we compared the obtained results of the proposed action recognition framework with the complete skeleton based RIAC-Net action descriptor results. And it is observed that part-wise RIAC-Net action descriptor performs superior.
- Finally, the effectiveness of the proposed framework for skeleton-based action recognition is shown, by achieving the state-of-the-art performance on three benchmark datasets such as UT Kinect Action 3D, Florence 3D actions Dataset and MSR Daily Action3D datasets.

Rest of the paper is organised as follows: - Section 2 discusses related work. In section 3 the details of the proposed work are illustrated. Datasets and the experiments are described in Section 4. Section 5 describes the experimental results. In the last, a conclusion is drawn about the proposed work and future work is discussed, in section 6.

2. Related Work

The skeleton-based action recognition approaches [20] [21] are progressing towards the temporal dynamics of the action using RNNs and LSTMs, recently. Du et al. [20] proposed an end-to-end hierarchical RNN to encode the relative motion between skeleton joints. Skeletons were split into anatomically relevant parts, which were fed into each independent subnet to extract local features. Shahroudy et al. [21] introduced a part-aware LSTM which possess part-based memory sub-cells and a new gating mechanism, showing the superior performance of LSTM over some hand-crafted features and RNN. To learn the human motion features of the skeleton sequence, RNN-LSTM [22] allows the network to access and store the long-range

contextual information of a skeleton sequence. Several authors [23] [24] exploited feature learning ability of CNNs which largely focused on a better skeletal representation and learning with simple CNNs. To better capture the Spatio-temporal dynamics of the skeleton sequences, some authors [25] [26] [27] used CNN as a spatial feature extractor and unified with RNN-LSTM network to model human motion. However, it is noticed that RNN-LSTM based approaches performed better. On the other side, the use of RNNs results in overfitting if number of input features are short enough to train the network and computational time dynamically increases with the number of input vectors.

Well-designed spatial-temporal encodings of skeleton sequences are more effective than only temporal information of skeleton-based action representation. Tu et al. [28] defined the correlation among three-dimensional signal using 3DCNN to capture spatial and temporal information of the action sequence. Liu et al. [29] mapped the skeleton joints in 3D coordinate space before extracting view-invariant Spatio-temporal features, which significantly improved the action recognition results. Whereas, the work [30] learnt adequate geometric features of 3D human actions by using Lie Group and unified it with deep neural networks. Chen et al. [31] encoded the skeleton joints as part based 5D feature vector to identify the most relevant joints of skeleton during the action sequence using a two-level hierarchical framework. Amor et al. [32] used trajectories on Kendall's shape manifolds to model the dynamics of human skeleton poses and used a parametrization-invariant metric for aligning, comparing, and modelling skeleton joint trajectories, to deal with noise caused by different execution rates of the actions performed by humans. However, this method is very time-consuming.

A good amount of work is also done to address the spatial representation of human skeleton poses which is mainly driven by the fact that an action can be characterized properly by the interactions or combinations of a subset of skeleton joints [22]. The methods to model action spatial patterns can be categorised in two classes: part-based model and sub-pose model. In the first category of spatial pattern modelling, the skeleton is divided into several groups, instead of considering the complete skeleton. The HBRNN [33] model decomposed the skeletons into five parts, two arms, two legs, and one torso, and built a hierarchical recurrent neural network to model the relationship among these parts. Similarly, Shahroudy et al. [21] proposed a part-aware LSTM model that construct the relationship between body parts. Whereas, in sub-pose model, the informative joints or their interactions are mainly focused. A handcraft features based approach [34] defined a SMIJ model which select the most informative joints by calculating statistical parameters such as mean and variance of joint angle trajectories and used the sequence of selected informative joints to represent the action. Wang et al. [35] mined co-occurrence distinctive spatial body-part structures as spatial part-sets and temporal evolutions of pose as temporal part sets. Whereas Lillo et al. [36] learnt the spatio-temporal annotations of complex

actions using motion poselets and actionlet dictionaries. Such annotations help to understand that which body part is active for a particular action but not discriminative enough in classification.

3. Proposed Work

In this section, the proposed framework for skeleton-based human action recognition is explained. It describes the formation of Compact Action Skeleton Sequence formation and proposed RIAC-Net architecture design in detail which includes three main steps- formation of Spatial-Temporal Convolution Features (STCF), defining Attention Driven Residual Block (ADRB), and lastly learning part-wise RIAC-Net based action features and to ensemble the predictions per part using weighted fusion scheme.

3.1 Compact Action Skeleton Sequence (CASS) generation

CASS is basically, a projection of each frame 3D coordinates of skeleton joints on the image frame which describes the spatial variation of the human skeleton pose during the action. The temporal details about the sequence of human poses are encrypted by using different colour coding for skeletons in such a way that colour of the skeletons change with time to exhibit the sequence of occurrence of frames. To exploit the discriminative local features of the actions, the generated CASS are further divided into five significant parts: i) Head to Spine (HS) ii) Left Leg (LL) iii) Right Leg (RL) iv) Left Hand (LH) and v) Right Hand (RH). Therefore, CASS is defined for FS as well as for other five parts of the skeleton - {HS, RL, LL, RH, LH}. Let an action video has n no. of frames $\{f_1, f_2 \dots\dots, f_n\}$, and each frame possess a human skeleton with k no. of skeleton joints $i.e.$ $\{(J_{x1},J_{y1},J_{z1}),(J_{x2},J_{y2},J_{z2}),\dots.(J_{xk},J_{yk},J_{zk})\}$, $k\epsilon[1,20]$. According to the configuration of skeleton joints in Figure 1, skeleton joints are partitioned in five groups as: $HS = \{J_4,J_3,J_2,J_1\}$, $LL = \{J_{13},J_{14},J_{15},J_{16}\}$, $RL = \{J_{17},J_{18},J_{19},J_{20}\}$, $LH = \{J_5,J_6,J_7,J_8\}$, $RH = \{J_9,J_{10},J_{11},J_{12}\}$. To generate CASS, the variations in joint coordinates of each group are sketched, mathematically defined as Eq. (1):

$$CASS_p = \begin{pmatrix} J_k^t & \cdots & J_k^n \\ \vdots & \ddots & \vdots \\ J_{k+4}^t & \cdots & J_{k+4}^n \end{pmatrix} \quad (1)$$

where k is the first skeleton coordinate number of each group, p is the partition label $\exists\ p\epsilon(HS, LL, RL, LH, RH)$ n is the number of frames in the action video.

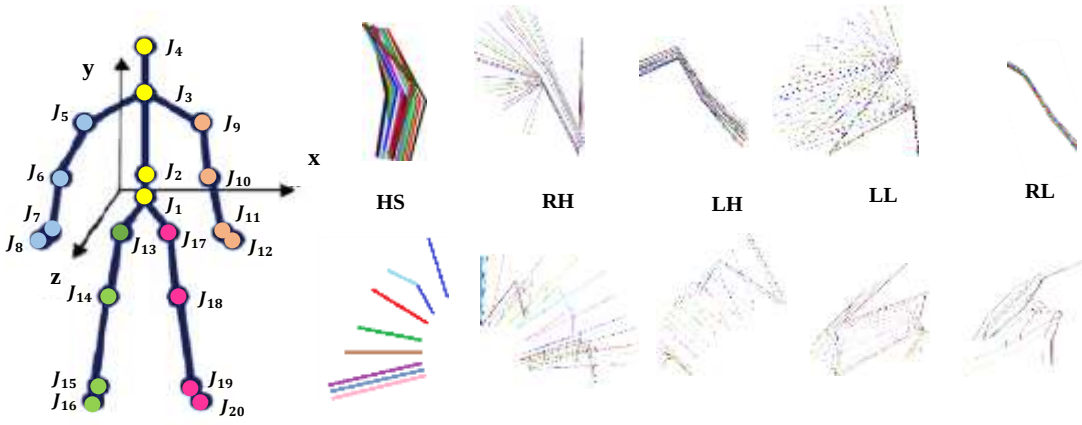


Fig. 1(a) Configuration of skeleton joints. Head to Spine (HS) joints is colored in Yellow, LH joints are colored in Blue, RH joints are coloured in Brown, LL joints are coloured in Green, RL joints are coloured in Pink (b) CASS formation for two actions – 'waving hand' and 'sitting down', of the Florence3D Action dataset

From the sample images of part-wise CASSs, for two actions- waving hand and sitting down, of the Florence3D Action dataset, in Fig. 1, it is observed that different amount of motion is associated with each part of the skeleton for an action resulting in unique patterns for each action. The part wise feature extraction and learning highlight the local dynamics of the skeleton. However, if the complete skeleton is processed to extract spatial deep features, the prominent movements in the action would be subsided.

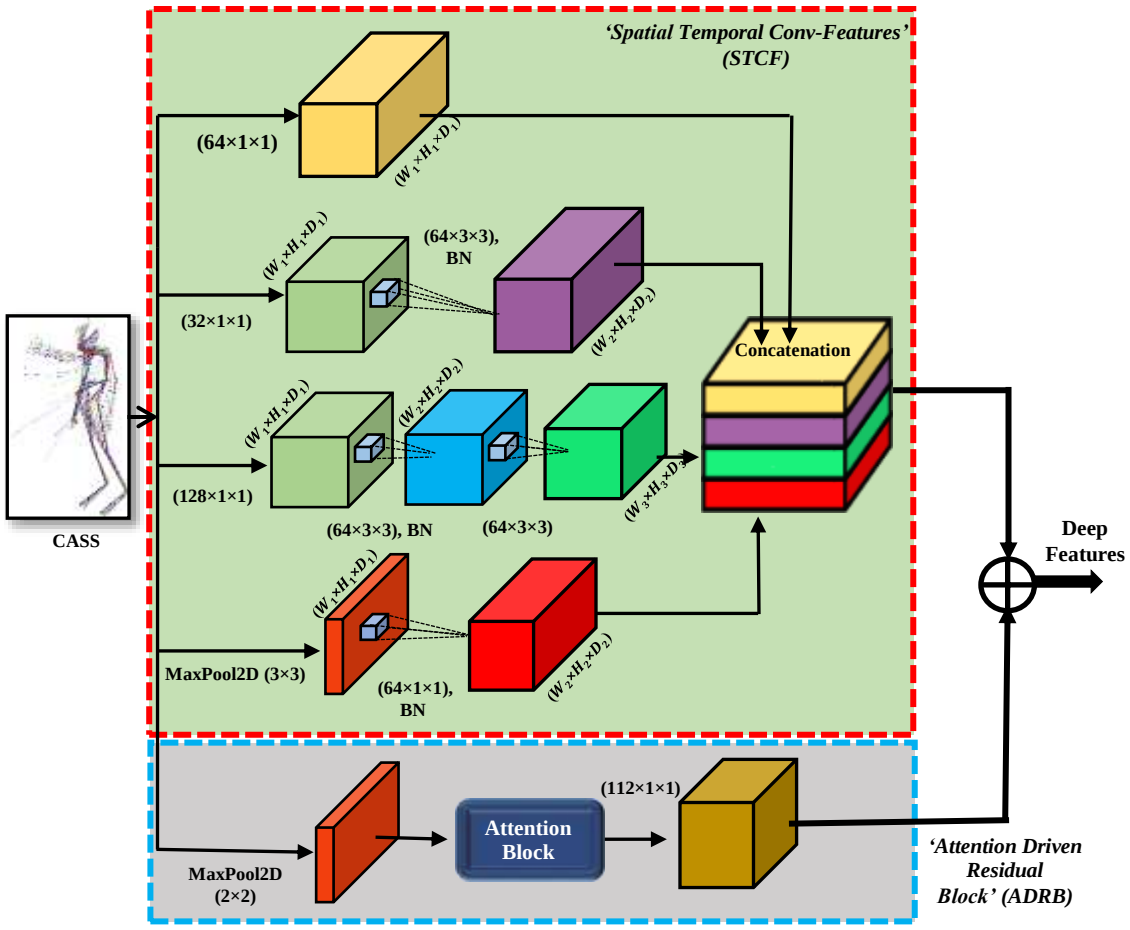


Fig. 2. Proposed Residual Inception Attention-based Convolution Network (RIAC-Net) block diagram

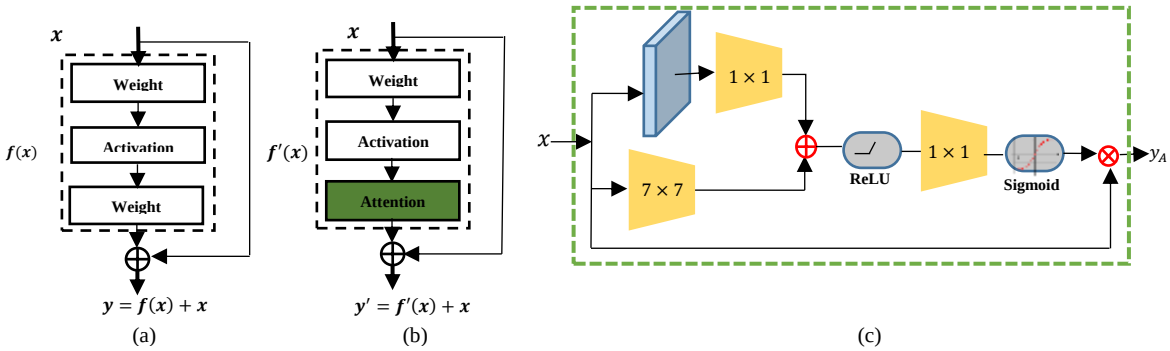


Fig. 3. Illustration of Attention Driven Residual Block architecture (a) Basic Residual Block (b) Attention Driven Residual Block (c) Attention block

3.2 Skeleton based action recognition with RIAC-Net

The training of Inception networks with residual connections has accelerated significantly, resulting in outperforming the similarly expensive inception networks without residual connections [17]. Therefore to solve the problem of skeleton-based action recognition for large inter-class similarity Residual Inception Attention-based Convolution Network (RIAC-Net), Fig. 2, is designed which is majorly divided into two parts- 'Spatial-Temporal Convolution Features' (STCF) and 'Attention Driven Residual Block' (ADRB).

A. Spatial-Temporal Convolution Features (STCF)

Salient features in an image, generally, can have an extremely large variation in size i.e. covering major section of the image or small section. Convolution helps to recover Spatio-temporal features only with right selection of kernel size. A large convolution kernel has large receptive field that highlights the globally distributed information and a smaller convolution kernel is preferred for locally distributed information. Use of multiple sized kernels in convolution filters i.e. $(1 \times 1, 3 \times 3, 5 \times 5)$ is an efficient solution to select the appropriate kernel size for good convolution features [18]. It essentially widens the network size and also computationally less expensive than deeper networks. It is the basic concept behind building the inception blocks [18] that targeted large size variations of spatial features. The convolution filters are made computationally more efficient by factorising (5×5) filters with two (3×3) filters in STCF. Description of the parameters of STCF block is provided in Table 1. It can be notified that equal-sized features $(W(112) \times H(112) \times D(64))$ are generated from all four branches of STCF block. And the final STCF feature vector is obtained by stacking STCF branch wise convolution features as $STCF_{FV} = \{STCFV_{FVi}\}$, where $i\epsilon[1,4]$ and each branch vector is constructed with dimension $(W_i \times H_i \times D_i) = (112 \times 112 \times 64)$ that results in $STCF_{FV}$ with $[448 \times 448 \times 64]$ dimesion.

B. Attention Driven Residual Block (ADRB)

The key structure of residual units allows direct signal propagation from first to the last layer of the network and gradients to propagate from the loss layer to any previous layer by skipping

the intermediate weight layers during backpropagation, which helps to handle the vanishing gradient problem to a great extent. Hence, the idea of residuals [35] proliferated the performance of deep networks by adding the identity function. The effect can be further enhanced by adding salient features instead of the identity function directly. It is implemented by using Attention Driven Residual Block (ADRB), shown in Fig. 3(a) and (b), where the attention block [36] tends to extract a spatial attention map by utilizing the inter-spatial relationship of features. The residual units are defined as follows:

$$y = \sigma_1\big(x + \mathcal{F}(x; \mathcal{W})\big) \tag{2}$$

where x and y are the input and output of the RIAC-Net architecture, $\sigma_1(.) \equiv$ ReLU [37], and $\mathcal{F}$ is a non-linear residual mapping function for input x which is formulated as follows:

$$\mathcal{F}(x; \mathcal{W}) = STCF_{FV}(x; \mathcal{W}_k) \tag{3}$$

where is $\mathcal{W} = \{\mathcal{W}_k, 4 \leq k \leq 1\}$ for each convolution branch of STCF block. The residual unit is modified as

$$y' = \sigma_1\big(\Psi(x) + \mathcal{F}(x; \mathcal{W})\big) \tag{4}$$

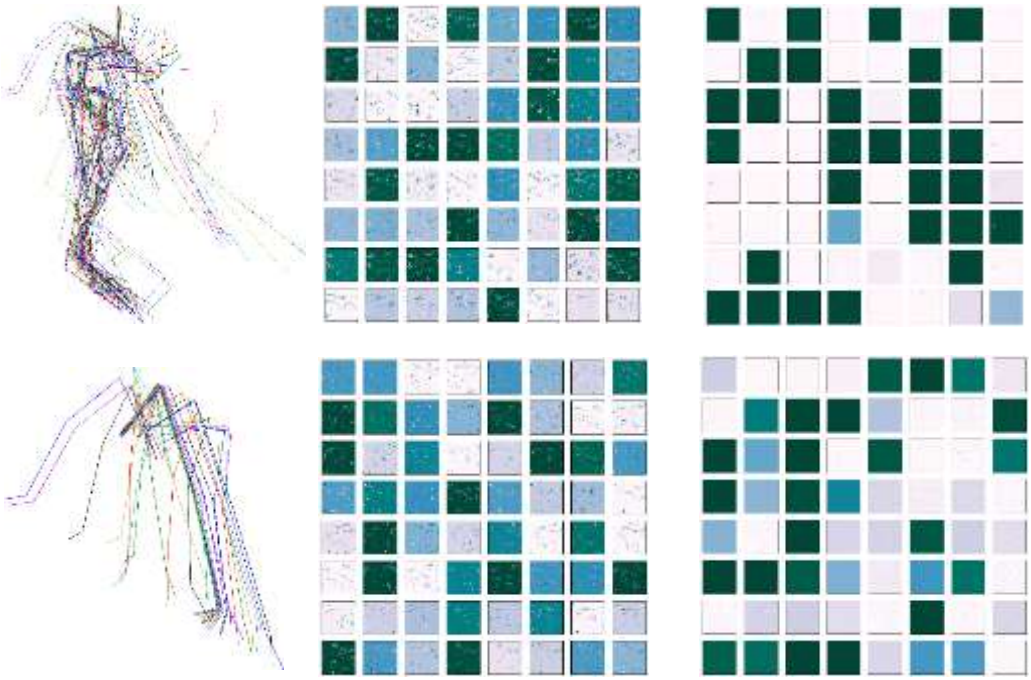
$$\text{And } \Psi(x) = x * \sigma_2\{f^{1\times1}\{\sigma_1\big[f^{7\times7}(x) + f^{1\times1}\big(\Lambda(x)\big)\big]\}\},\ \Psi\epsilon\mathbb{R}^{W\times H} \tag{5}$$

where $\sigma_1$is a ReLU function and σ_2 is a sigmoid function, $f^{1\times1}(.), f^{7\times7}(.)$ are non-linear convolution layers with 1×1 and 7×7 kernel sizes and $\Lambda(.)$ is a 2D max-pool layer with 2×2 kernel size. The '*' and '+' are multiplicative [38] and additive [39] attention operator. To extract salient features additive attention method performs better for large dimensional input features [40] whereas the multiplicative attention method holds fast computations and also memory-efficient due to the matrix multiplication. Therefore, at the input stage additive attention operator is applied to handle larger input dimension than the later one, as shown in Fig. 3 (c). The two different sized kernels in the convolution layer $f^{k\times k}$ cover the features on the coarse spatial grid level and finer grid level which collectively helps to identify relevant features and disambiguate the task irrelevant features in x.

Table 1: Description RIAC-Net –STCF Block Architecture parameters

RIAC-Net Branches	**No. of filters**	**Kernel size/stride**	**Input size** $(W_I \times H_I \times D_I)$	**Output size** $(W_O \times H_O \times D_O)$
Branch 1	Conv, 64	$(1 \times 1)/2$	$(224 \times 224 \times 3)$	$(112 \times 112 \times 64)$
Branch 2	Conv, 32	$(1 \times 1)/1$	$(224 \times 224 \times 64)$	$(224 \times 224 \times 32)$
	Conv, 64	$(3 \times 3)/2$	$(224 \times 224 \times 32)$	$(112 \times 112 \times 64)$
Branch 3	Conv, 128	$(1 \times 1)/1$	$(224 \times 224 \times 3)$	$(224 \times 224 \times 128)$
	Conv, 64	$(3 \times 3)/1$	$(224 \times 224 \times 128)$	$(224 \times 224 \times 64)$
	Conv, 64	$(3 \times 3)/2$	$(224 \times 224 \times 64)$	$(112 \times 112 \times 64)$
Branch 4	$Maxpool$	$(2 \times 2)/1$	$(224 \times 224 \times 3)$	$(112 \times 112 \times 3)$
	Conv, 64	$(1 \times 1)/1$	$(112 \times 112 \times 3)$	$(112 \times 112 \times 64)$

Sample images of the residue and attention driven residue activation maps of the action, for complete CASS and part wise CASS, are shown in Fig. 4. It is clearly observed that the residue branch extracts noisy data whereas attention driven residue branch highlight the salient information about the action resulting in improved recognition performance.



(a) (b) (c)

Fig. 4. Illustration of Residue and Attention driven residue activation maps (8×8) for complete CASS and left leg (LL) CASS (a) input CASS (b) residue activation maps (c) attention driven residue activation map skeleton human action recognition.

C. Learning of part-wise RIAC-Net based action descriptors

Late fusion approach works better than the early fusion scheme [41] at the cost of additional learning attempts. Therefore, RIAC-Net based action descriptor is designed and learnt for each part-wise CASS, individually, using the combination of global average pooling (GAP), batch normalisation (BN), Long Short Term Memory (LSTM) layers, dropout layer with 0.2 dropout probability and dense layer. The final prediction is given by fusing the learnt part-wise CASS based predictions using weighted fusion scheme, as shown in Fig. 5. The best recognition performance for a specific weight combination i.e. $\{w_i\}\exists i \in [1,5]$, is reported finally. To learn the unique patterns of the part wise RIAC-Net based action features two consecutive LSTM layers are used in such a way that output gate $\boldsymbol{h_t^2}$ of former LSTM layer is fed to the input gate $\boldsymbol{i_t^2}$ of later LSTM layer. Let x_t^i , h_t^i , and C_t^i be input, output and cell state of the i^{th} LSTM layer, at instance t respectively. The sequence of flow of the signal from the first LSTM layer to second layer is given by Eq. 6 to 9.

$$h_t^1 = \sigma(W_0^1 * [h_{t-1}^1, x_t^1] + b_o) * \tanh(C_t^1) \quad (6)$$

$$C_t^1 = C_{t-1}^1 * f_t^1 + \widehat{C_t^1} * i_t^1 \quad (7)$$

$$x_t^2 = h_t^1 \; and \; h_t^2 = \sigma(W_0^2 * [h_{t-1}^2, x_t^2] + b_o) * \tanh(C_t^2) \quad (8)$$

$$C_t^2 = C_{t-1}^2 * f_t^2 + \widehat{C_t^2} * i_t^2 \quad (9)$$

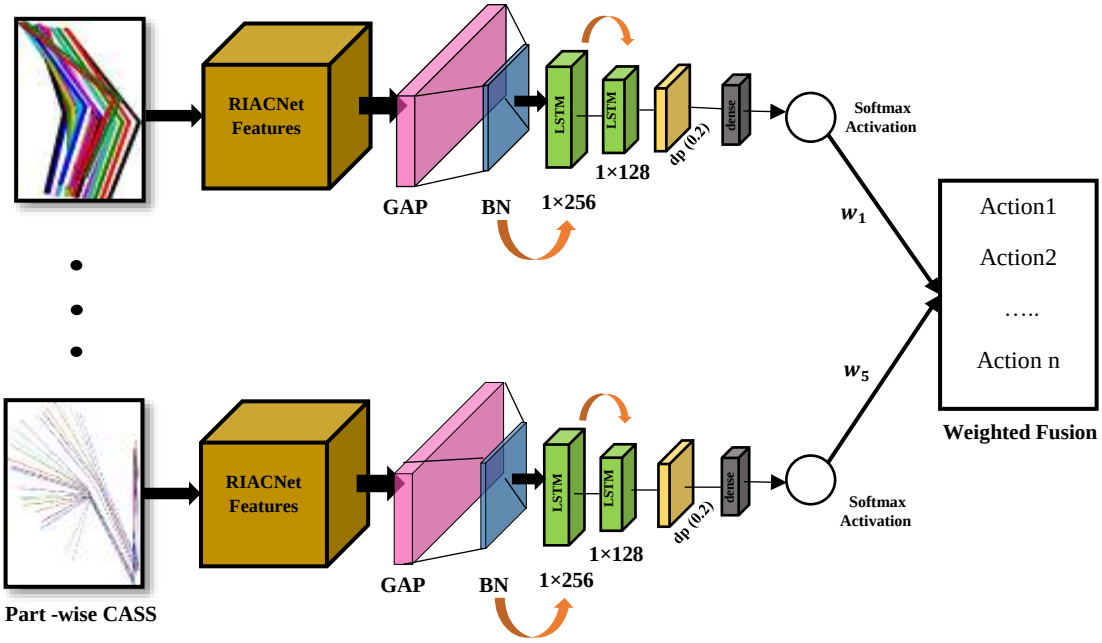


Fig. 5. Description of proposed Part-wise Spatiotemporal and Attention Driven Residues based Learning for skeleton human action recognition.

where σ is the sigmoid operation, W_0^i is the weights of the output gate of the i^{th} LSTM layer, cell state $C_t^i$of the i^{th} layer is computed using both forget gate output as f_t^i and input gate output as i_t^t. $\widehat{C_t^i}$ control the amount of update required to current cell state C_t^i, according to the input x_t^i passed through input gate, i_t^i. The learnt vector h_t^2 with $[1 \times 128]$ dimension is fed to dropout layer to handle the overfitting problem followed by a dense layer and Softmax activation function. Predictions from each part-wise branch are fused using weighted fusion that utilises all possible combination of weights w_i to find the best set of weights, as shown follows:

$$P_c = \sum_{i=1}^{5} w_i \times p_i \, , c \in (1, n) \quad (10)$$

where n is the number of classes.

4. Experiments and Results

To validate the performance of the proposed framework for skeleton-based human action recognition, three publically available 3D datasets- UT Kinect Action 3D and Florence 3D actions Dataset and MSR Daily Action3D datasets.

4.1 UT Kinect Action 3D dataset

UT Kinect Action 3D dataset [42] captures 10 types of human actions in indoor settings, using single stationary Kinect hardware. The dataset includes an RGB image with 640×480 resolution, depth image with 320×240 and twenty 3D joints of a human skeleton per frame captured at 30 FPS. The 10 actions include i) walk, ii) sit down iii) stand up iv) pick up v) carry vi) throw vii) push viii) pull ix) wave and x) clap hands. These actions are performed by 10 different actors (9 males and 1 female), each repeated two times. Hence, it consists of a total 200 (10×10×2) action samples with 6220 frames. The challenge lies in the fact that there exist viewpoint variations and high intra-class variations. The length of the sample actions ranges from 5 frames to 120 frames. Therefore, in the proposed work the number of frames for each action sample is made equal to 60 frames, before generating CASS representation of the action. It is performed by down-sampling the frames of the actions which possess more than 60 frames. And up-sampling is applied to the actions which possess less than 60 frames to maintain a symmetricity in the CASS generated using action frames. The sample RGB images are shown in Fig. 6 below. We use the skeleton representation of human actions for human action recognition.



Fig. 6. Sample images of the publically available UT-Kinect Action 3D Dataset

4.2 Florence 3D Action dataset

Florence 3D Action dataset [43] dataset is collected at the University of Florence using a Kinect camera. It includes 9 actions: i) arm wave, ii) drink from a bottle, iii) answer phone, iv) clap, v) tight lace, vi) sit down, vii) stand up, viii) read watch, ix) bow, as shown in Fig. 7. Each action is performed by 10 subjects several times for a total of 215 sequences. The sequences are acquired using the OpenNI SDK, with skeletons represented by 15 joints instead of 20 as with the Microsoft Kinect SDK. The main challenges of this dataset are the similarity between actions,

the human object interaction, and the different ways of performing the same action.

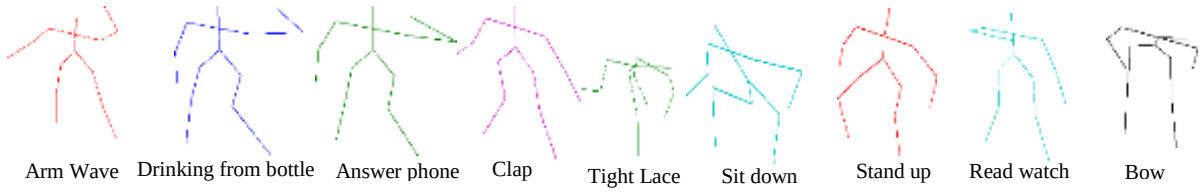


Fig. 7. Sample frames of Florence Action 3D dataset

4.3 MSR Action 3D Dataset

The MSR Action 3D dataset [44] consists of 20 different action classes. Each action is performed by 10 subjects for three times. It is divided into 3 subsets AS1, AS2, AS3 each include 8 actions, as shown in Table 2. It includes 567 skeleton sequences. However, skeletons for 10 sequences are missing. Therefore, the experiments are conducted on 557 sequences, each provided with 3D coordinates of 20 joints per frame. For experimentation, cross-subject evaluation protocol is used. According to which, half of the dataset with 1, 3, 5, 7 and 9 subject ids, is used for training and another half of the dataset with 2, 4, 6, 8, 10 subject ids is used for testing for each action set-AS1, AS2, AS3.

Table 2: List of action classes in each subset AS1, AS2, and AS3 of the MSR Action 3D dataset

Activity set 1 (AS1)	Horizontal arm wave	Hammer	Forward punch	High Throw	Hand clap	Bend	Tennis serve	Pick up and throw
Activity set 2 (AS2)	Horizontal arm wave	Hand catch	Draw X	Draw Tick	Draw Circle	Two hand wave	Forward kick	Side boxing
Activity set 3 (AS3)	High throw	Forward kick	Side kick	Jogging	Tennis Swing	Tennis Serve	Golf Swing	Pick up and Throw

Data augmentation: Deep neural networks demand a large amount of training data to perform efficiently. We have only 557,428, and 200 skeleton sequences, for MSR Action 3D dataset, Florence 3D Action dataset and UT Kinect Action 3D Dataset. Thus, to prevent overfitting, data augmentation techniques are applied before processing the data. It includes cropping, horizontal and vertical flip, rotation at 45^0 and -45^0.

The RIAC-Net architecture is implemented and evaluated in Python with Keras framework using the Tensor-Flow backend. We used mini-batches of 256 images, and Adam optimizer with default parameters, $\beta1 = 0.9$ and $\beta2 = 0.999$, during training. The initial learning rate is set to 0.001 and is decreased by a factor of 0.02 after every 20 epochs. The network is trained for each set of part wise input (HS, LH, RH, LL, RL), for 1000 epochs from scratch. To handle the overfitting in the training phase, we adopted weight noise and early stopping [45] along with drop-out strategy.

Table 3: Performance of the proposed framework for UT Kinect dataset

	Full skeleton	Head Spinal	Left Leg	Right Leg	Left hand	Right hand	Weighted Fusion
Training Loss	0.2836	0.4480	0.4697	0.3209	0.3164	0.3801	$W_{HS}, W_{LL}, W_{RL}, W_{LH}, W_{RH}$
Training Accuracy	99.96	92.94	92.94	97.65	97.13	95.92	{2,3,4,4,5}
Test accuracy	97.71	97.49	97.49	96.45	95.94	96.48	**100.00**

Table 4: Performance of the proposed framework for Florence 3D Action Dataset

	Full skeleton	Head Spinal	Left Leg	Right Leg	Left hand	Right hand	Weighted Fusion
Training Loss	0.1221	0.3700	0.0279	0.0630	0.0868	0.0127	$W_{HS}, W_{LL}, W_{RL}, W_{LH}, W_{RH}$ {3,4,2,3,2}
Training Accuracy	100.00	100.00	100.00	99.69	96.65	100.00	
Test accuracy	95.89	100.00	92.47	95.89	92.63	91.85	**98.33**

The experimental results on UT Kinect 3D Action dataset, Florence 3D Action dataset and MSR Action 3D dataset are reported in Table 3, 4 and 5 respectively. The weights $w_i, i\epsilon(1,5)$ corresponding with the best test accuracy achieved are also provided in Table 3, 4 and 5. The Validation loss curves for UT Kinect 3D Action dataset, Florence 3D Action dataset and MSR Action 3D dataset are shown in Fig. 7 (a), (b), and (c)-(e), respectively.

Table 5: Performance of the proposed framework for MSR action 3d dataset under cross-subject evaluation strategy

Dataset subsets	Full skeleton	Head Spinal	Left Leg	Right Leg	Left hand	Right hand	Weighted Fusion Accuracy $W_{HS}, W_{LL}, W_{RL}, W_{LH}, W_{RH}$
AS1	94.8	96	95	98.7	96	90	96.7 { **2,3,3,3,3** }
AS2	93.3	97.33	100.00	99.5	96.44	93.3	**98.6 { 1,5,4,1,5 }**
AS3	96.5	100	98.2	100	97.3	95.4	**99.9 { 2,4,1,1,3 }**
Overall	94.8	97.77	96.84	99.39	96.58	92.9	**98.40**

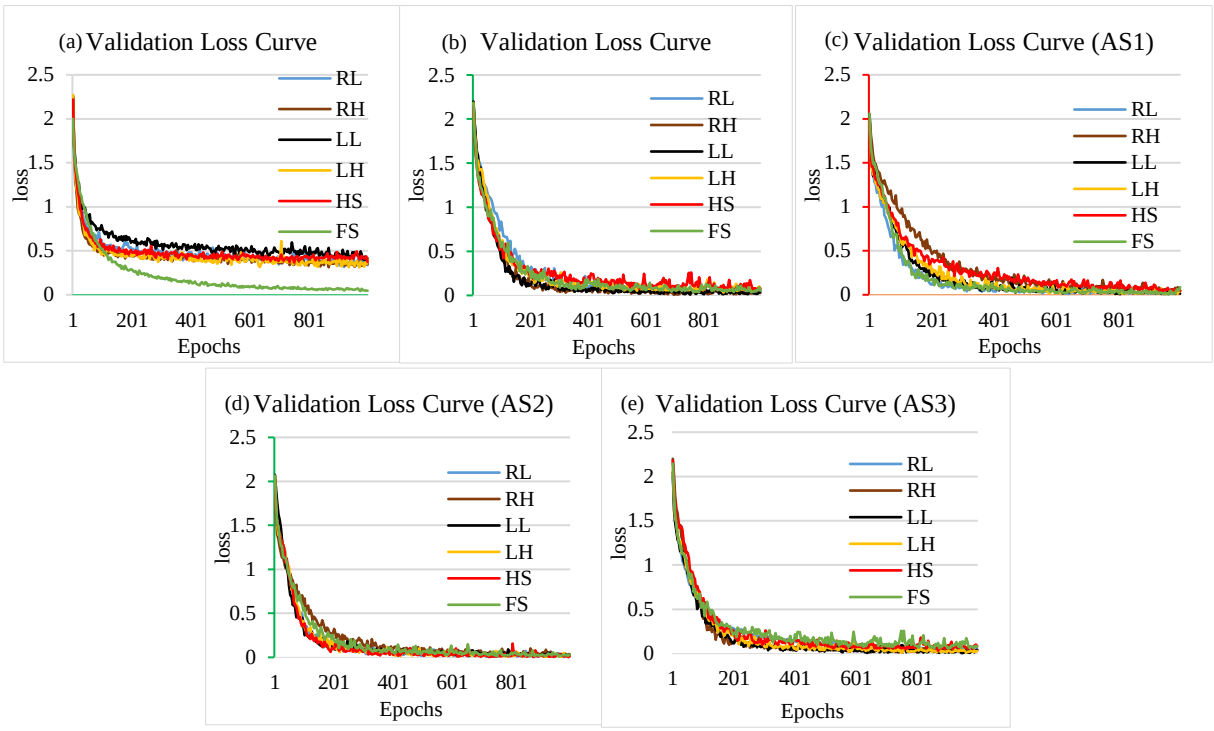


Fig.7. Illustration of Part-Wise (RL, RH, LL , LH, HS) and full skeleton (FS) based validation loss curves for (a) UT Kinect 3D Action Dataset, (b) Florence 3D Action dataset, (c)-(e) MSR Action 3D dataset AS1, AS2, and AS3

The Validation losses gradually decrease with the epochs, which confirms the adequate learning of the models for each part. Receiver output Characteristic (ROC) curve plotted between True Positive Rate (TPR) and False Positive Rate (FPR), in Fig.8 (a) and (b), also support the fact that weighted fusion of part-wise skeleton RIAFNet features turned out better action representation than Full skeletons (FS) for UT Kinect 3D Action dataset, Florence 3D Action dataset and MSR Action 3D dataset. Area Under the Curve (AUC), i.e. $AUC \in (0,1)$, is also computed for each method. The highest AUC values 1.00, 0.97, (0.97,0.99, and 1.00) are

obtained for weighted average late fusion strategy over FS and part-wise (HS, LL, LH, RH, RL) skeleton based approaches for UT Kinect 3D Action dataset, Florence 3D Action dataset, MSR Action 3D dataset-AS1, AS2, AS3 sets respectively.

Table 6: A comparison of the proposed framework with other state-of-the-arts for UT Kinect Action dataset

Method	Learning Model	Protocol	Accuracy (%)
Feature Combination [46]	K-NN	LOOCV	98
ST-LSTM+Trust Gate [20]	Hierarchal RNN	LOOCV	97
Grassmann Manifold [47]	LTBSVM	LOOCV	88.5
Geometric Features [11]	Multi-layer LSTM	cross validation	95.96
TS-LSTM [48]	Ensemble Temporal sliding LSTM	cross validation	96.97
Lie groups [7]	SVM	One vs all cross validation	97.08
Kernel Linearization [49]	SVM	cross validation	98.2
LRCNLG [30]	LSTM	LOCCV	98.5
Proposed work	**LSTM layers**	**LOOCV**	100.00

Table 7: A comparison of the proposed framework with other state-of-the-arts for Florence 3D Action dataset

Method	Learning Model	Protocol	Accuracy (%)
Lie groups [7]	SVM	One vs all Cross validation	90.88
Kernel Linearization [49]	SVM	Cross validation	95.23
Riemannian Manifold [50]	K-NN	LOOCV	87.04
Mining key pose [51]	Inference Algorithm	LOOCV	92.25
Feature combination [46]	K-NN	LOOCV	94.39
LRCNLG [30]	LSTM	LOOCV	95.37
Proposed work	**LSTM layers**	**LOOCV**	**98.33**

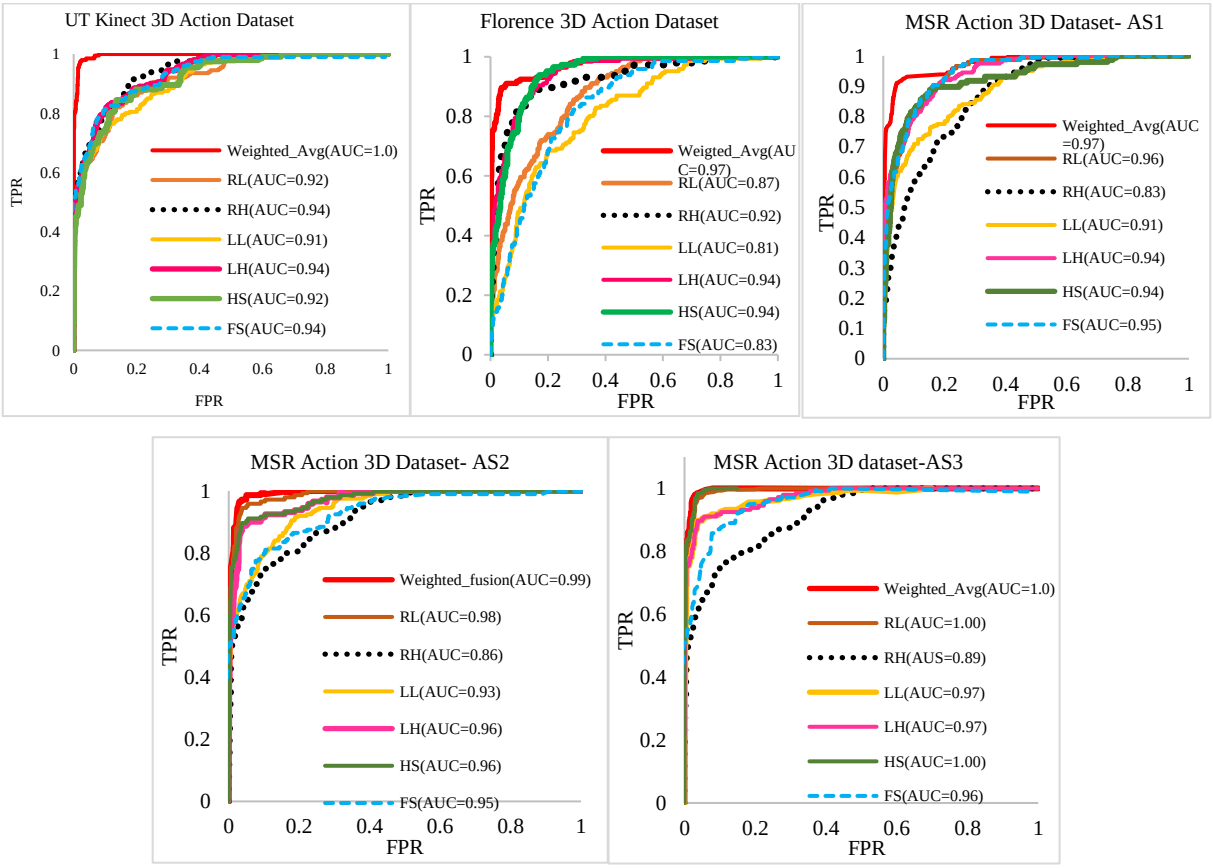


Fig. 8. ROC Curves of a) UT Kinect 3D dataset b) Florence Action 3D dataset c) MSR Action 3D dataset.

The achieved accuracy of the proposed work is compared with the other-state-of-arts for UT Kinect 3D Action dataset, Florence 3D Action dataset and MSR Action 3D dataset in Table 6, 7, and 8. The obtained results outperforms many previous studies [7]- [30], [10] - [14]. The

proposed work achieved 100% recognition accuracy for UT Kinect 3D Action dataset using Leave-One-Out Cross Validation (LOOCV) scheme.

Table 8: A comparison of the proposed framework with other state-of-the-arts for MSR Action 3D dataset cross subject evaluation

Method	Learning Model	AS1	AS2	AS3	Average Accuracy (%)
HBRNN-L [10]	Hierarchical RNN	93.33	94.64	95.50	94.49
ST-NBNN [33]	Naive-Bayes Nearest-Neighbor	91.50	95.60	97.30	94.80
DMM-LBP-DF [52]	K-ELM	98.10	92.00	94.60	94.90
VBDDM [53]	ProCRC	99.10	92.30	98.20	96.50
Mean3DJ [16]	Random Forest	-	-	-	82.68
Lie group-MinP-PrefixSpan [54]	SVM	-	-	-	97.4
SPMFs [15]	D-CNN	97.06	99.00	98.09	98.05
ResNet-44 [14]	D-CNN	99.90	99.80	100	99.90
Proposed method	**LSTM layers**	**95.70**	**98.60**	**99.90**	**98.06**

The Classification result for each AS1, AS2, and AS3 sets of the MSR-Action3D dataset are shown in Fig. 9 (c), (d), and (e). It is noticed that misclassification occurs only for the actions with high inter class similarity such as 'draw tick' and 'draw *X*', 'Pickup and Throw' and 'Bend'. Whereas "Forward Kick" and "Tennis Serve" actions which share a large overlap in the sequences, are more challenging to distinguish the two actions in AS3 set. The proposed framework handled this inter class similarity between the two actions and recognised 'Forward Kick' and 'Tennis Serve' with 100% accuracy.

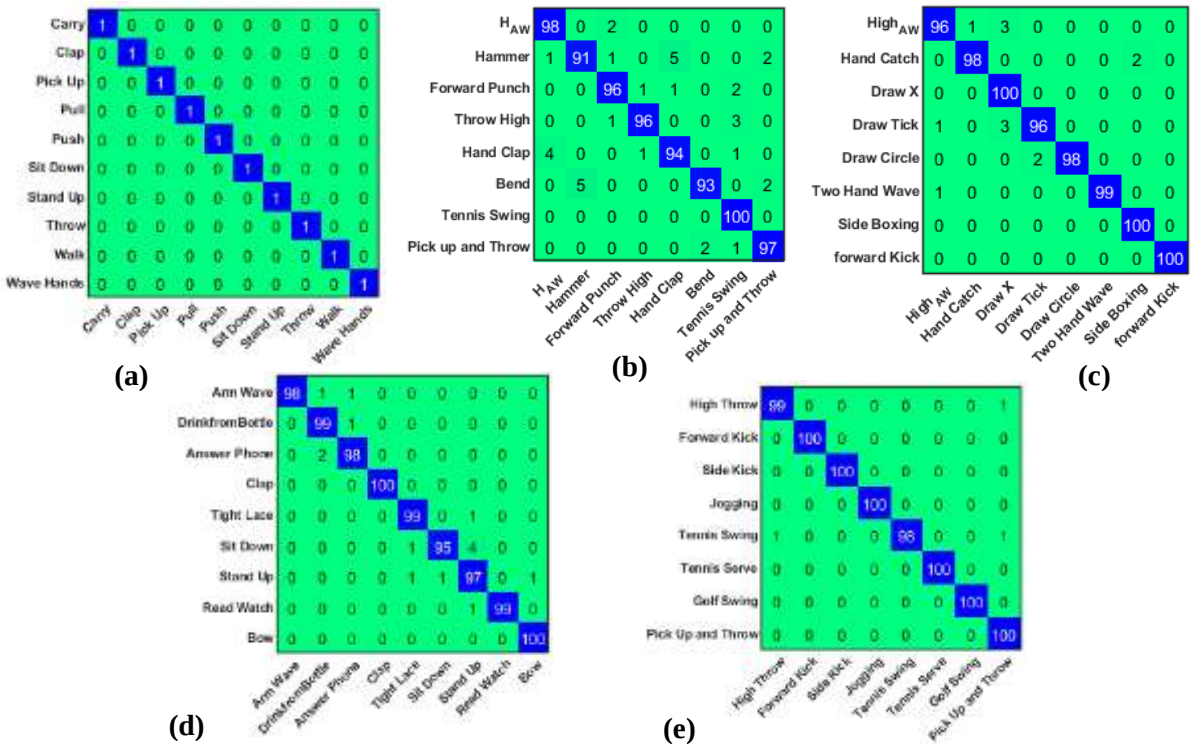


Fig. 9. Confusion Matrix of the a) UT Kinect Dataset, b) Florence Action 3D dataset and MSR action 3D dataset c) AS1, d) AS2 and e) AS3 sets. Where H_A_W: Horizontal Arm Wave and High_A_W: High Arm Wave

The weighted classification confusion matrix of the UT Kinect 3D action dataset is shown in Fig. 9 (a). From where it is clearly evident that each action is recognized correctly without any misclassification irrespective of the presence of high intra class variations and view variations.

The obtained result outperforms many previous studies Lie groups [7] , LRCNLG [30] , Grassmann Manifold [47], TS-LSTM [48] which tried to learn geometrical 3D features of human actions using Lie groups, Grassmann Manifold and temporal sliding LSTMs respectively.

The proposed work achieved 98.33% recognition accuracy on Florence 3D Action dataset, which is 2.96% higher than LRCNLG [30] that integrated Lie groups with deep neural networks to learn the geometrical 3D features. The confusion matrix for Florence 3D action dataset is shown in Fig. 9 (b) which shows that the proposed work obtained very high accuracy for most of the actions. However, there exist some confusion between similar actions such as 'Answer Phone', 'drink from bottle', and 'high arm wave', 'stand up' and 'sit down'. We have achieved 98.05% fairly a high recognition accuracy on MSR action 3D dataset which outperformed previous works [15] [33]. However, Pham et al. [14] achieved 99.90% accuracy which utilised deep ResNets to process skeleton data for human action recognition. Some skeleton based methods like [33] [15] used skeleton features based on pairwise distances between joints. However, our results obtained on MSR action 3-D dataset show that part wise analysis of whole skeletons followed by late fusion approach is more discriminative approach than taking into consideration the joints separately.

5. Conclusion

In this paper, an effective skeleton based part-wise spatio-temporal CNN - RIAFNetwork based 3D human action recognition framework is proposed. It models the dynamics of the action by splitting the skeletons into five parts- Head to Spine (HS), Left Leg (LL), Right Leg (RL), Left Hand (LH), Right Hand (RH). Each part of the skeleton behaves differently for every action which is encrypted using RIAC-Net network which helps to highlight local dynamics {LL, LH, RH, RL, HS} of the action, that proved superior representation than the global action dynamics {FS} of the skeleton. The architecture of the RAIC-Net is designed using the concept of attention based residues and inception blocks. The final action class scores are generated by weighted (decision level) fusion of deep features. The empirical results and the analysis of the performance of our proposed approach exhibit promising results with high accuracies 100%,98.03%, and 98.7% on UT Kinect Action 3D and Florence 3D actions Dataset and MSR Daily Action3D datasets. Obtained results show that weighted fusion of part wise skeleton action dynamics' learning performs better than FS based action recognition. It is also observed that the proposed model is able to handle the intra class variations and inter class similarity among the actions quite decently.

In future, the work can be extended to be capable enough to handle view variations and partial occlusions in real-time scenario, which are the two very challenging aspects for a robust action recognition system.

References

[1] C. Dhiman and D. K. Vishwakarma, “A review of state-of-the-art techniques for abnormal human activity recognition,” *Engineering Applications of Artificial Intelligence,* vol. 77, pp. 21-45, 2019.

[2] J. K. Aggarwal and L. Xia, “Human activity recognition from 3D data: A review,” *Pattern Recognition Letters,* vol. 48, pp. 70-80, 2014.

[3] K. Singh, S. Rajora, D. K. Vishwakarma, G. Trapathi, S. Kumar and G. S. Walia, “Crowd anomaly detection using Aggregation of Ensembles of fine-tuned ConvNets,” *Neurocomputing,* 2019.

[4] D. Ludl, T. Gulde and C. Curio, “Simple yet efficient real-time pose-based action recognition,” *arXiv:1904.09140v1 [cs.CV],* 2019.

[5] M. Straka, S. Hauswiesner, M. Rüther and H. Bischof, “Skeletal Graph Based Human Pose Estimation in Real-Time,” in *British Machine Vision Conference (BMVC),* Dundee, UK, 2011.

[6] K. Li and Y. Fu, “Prediction of human activity by discovering temporal sequence patterns,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 36, no. 8, pp. 1644-1657, 2014.

[7] R. Vemulapalli , F. Arrate and R. Chellappa, “Human action recognition by representing 3d skeletons as points in a lie group,” in *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 2014.

[8] J. Wang, Z. Liu, Y. Wu and J. Yuan, “Mining actionlet ensemble for action recognition with depth cameras,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Providence, Rhode Island, 2012.

[9] D. Wu and L. Shao, “Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Columbus, Ohio, 2014.

[10] Y. Du, Y. Fu and L. Wang, “Representation Learning of Temporal Dynamics for Skeleton-Based Action Recognition,” *IEEE Transactions on Image Processing,* vol. 25, no. 7, pp. 3010-3022, 2016.

[11] S. Zhang , X. Liu and J. Xiao, “On Geometric Features for Skeleton-Based Action Recognition Using Multilayer LSTM Networks,” in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, Santa Rosa, CA, USA, 2017.

[12] D. K. Vishwakarma and R. Kapoor, “Integrated Approach for Human Action Recognition using Edge Spatial Distribution, Direction Pixel, and R -Transform", Advanced Robotics,” *Advanced Robotics,* vol. 29, no. 23, pp. 1551-1561, 2015.

[13] D. K. Vishwakarma and K. Singh, “Human Activity Recognition based on Spatial Distribution of Gradients at Sub-levels of Average Energy Silhouette Images,” *IEEE Transactions on Cognitive and Developmental Systems,* vol. 9, no. 4, pp. 316-327, 2016.

[14] H. H. Pham, L. Khoudour, A. Crouzil and P. Zegers, “Exploiting deep residual networks for human action recognition from skeletal data,” *Computer Vision and Image Understanding,* 2018.

[15] H. H. Pham, L. Khoudoury, A. Crouzil, P. Zegers and S. A. Velastiny, “Skeletal movement to color map: A novel representation for 3D Action Recognition with Inception Residual Networks,” *arXiv:1807.07033v1 [cs.CV] ,* 2018.

[16] A. Tamou , L. Ballihi and D. Aboutajdine , “Automatic Learning of Articulated Skeletons based on Mean of 3D Joints for Efficient Action Recognition,” *International Journal of Pattern Recognition and Artificial Intelligence,* vol. 31, no. 4, pp. 1-17, 2017.

[17] C. Szegedy, S. Ioffe and V. Vanhoucke, “Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning,” *arXiv:1602.07261v2 [cs.CV],* 2016.

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich, "Going deeper with convolutions," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 2015.

[19] K. He , X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," *arXiv:1512.03385v1 [cs.CV]*, 2015.

[20] Y. Du, W. Wang and L. Wang, "Hierarchical Recurrent Neural Network for Skeleton Based Action Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, 2015.

[21] A. Shahroudy, J. Liu, T.-T. Ng and G. Wang, "NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016.

[22] W. Zhu , C. Lan , J. Xing, W. Zen, Y. Li, L. Shen and X. Xie, "Cooccurrence Feature Learning for Skeleton based Action Recognition using Regularized Deep LSTM Networks," in *30th AAAI Conference on Artificial Intelligence (AAAI)*, Phoenix, Arizona USA, 2016.

[23] Y. Hou , Z. Li, P. Wang and W. Li , "skeleton Optical Spectra Based Action Recognition Using Convolutional Neural Networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 3, pp. 807-811, 2018.

[24] S. Song', C. Lan , J. Xing , W. Zeng and J. Liu , "An End-to-End SpatioTemporal Attention Model for Human Action Recognition from Skeleton data," in *Thirty-First AAAI Conference on Artificial Intelligence*, California, USA, 2017.

[25] T. S. Kim and A. Reiter, "Interpretable 3D Human Action Analysis with Temporal Convolutional Networks," *arXiv preprint arXiv:1704.04516*, 2017.

[26] Z. Shi and T.-K. Kim, "Learning and Refining of Privileged Information-based RNNs for Action," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, Hawaii, 2017.

[27] B. Mahasseni and S. Todorovic, "Regularizing Long Short Term Memory with 3D Human-Skeleton Sequences for Action Recognition," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, United States, 2017.

[28] J. Tu, M. Liu and H. Liu, "Skeleton based Human Action Recognition Using Spatial Temporal 3D Convolutional Neural Network," in *IEEE International Conference on Multimedia and Expo*, San Diego, CA, USA, 2018.

[29] M. Liu , H. Liu and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognition*, vol. 68, pp. 346-362, 2017.

[30] M. Rhif, H. Wannous and I. R. Farah, "Action Recognition from 3D Skeleton Sequences using Deep Networks on Lie Group Features," in *International Conference on Pattern Recognition (ICPR)*, Beijing, China, 2018.

[31] H. Chen, G. Wang, J.-H. Xue and L. He, "A novel hierarchical framework for human action recognition," *Pattern Recognition*, vol. 55, pp. 148-159, 2016.

[32] B. B. Amor, J. Su and A. Srivastava, "Action recognition using rate-invariant analysis of skeletal shape trajectories," *IEEE Transaction on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 38, no. 1, pp. 1-13, 2016.

[33] J. Weng , C. Weng and J. Yuan, "SpatioTemporal Naive-Bayes Nearest-Neighbor (ST-NBNN) for skeleton-based action recognition," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, Hawaii, 2017.

[34] F. Ofli , R. Chaudhry , G. Kurillo , R. Vidal and R. Bajcsy, "Sequence of the Most Informative Joints (SMIJ): A new representation for human skeletal action recognition," *Journal of Visual Communication and Image Representation*, vol. 25, no. 1, pp. 24-38, 2014.

[35] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, "Attention Is All You Need," in *31st Conference on Neural Information Processing Systems (NIPS)*, Long Beach, CA, USA, 2017.

[37] K. He , X. Zhang, S. Ren and J. Sun, "Identity mappings in deep residual networks," *arXiv preprint arXiv:1603.05027*, 2016.

[38] E. Shelhamer, J. Long and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence* , vol. 39, no. 4, pp. 640-651, 2017.

[39] D. Bahdanau, K. Cho and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv:1409.0473*, 2014.

[40] D. Britz, A. Goldie, M. T. Luong and Q. Le, "Massive exploration of neural machine translation architectures," *arXiv:1703.03906*, 2017.

[41] C. G. Snoek, M. Worring and A. W. Smeulders, "Early versus Late Fusion in Semantic Video Analysis," in *Thirteenth ACM International Conference on Multimedia*, Singapore, 2005.

[42] L. Xia , C. C. Chen and J. K. Aggarwal, "View invariant human action recognition using histograms of 3D joints," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Providence, RI, USA, 2012.

[43] L. Seidenari , V. Varano , S. Berretti , A. D. Bimbo and P. Pala, "Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Portland, OR, USA, 2013.

[44] W. Li , Z. Zhang and Z. Liu, "Action recognition based on a bag of 3d points," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, San Francisco, CA, USA, 2010.

[45] A. Graves, "Practical variational inference for neural networks," in *International Conference on Neural Information Processing Systems*, Granada, Spain , 2011.

[46] D. C. Luvizon, . H. Tabia and D. Picard, "Learning features combination for human action recognition from skeleton sequences," *Pattern Recognition Letters*, vol. 99, no. 1, pp. 13-20, 2017.

[47] R. Slama, H. Wannous, M. Daoudi and A. Srivastava, "Accurate 3D action recognition using learning on the Grassmann manifold," *Pattern Recognition*, vol. 48, no. 2, pp. 556-567, 2015.

[48] I. Lee, D. Kim , S. Kang and S. Lee, "Ensemble Deep Learning for Skeleton-Based Action Recognition Using Temporal Sliding LSTM Networks," in *IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017.

[49] P. Koniusz, A. Cherian and F. Porikli, "Tensor representations via kernel linearization for action recognition from 3d skeletons," in *European Conference on Computer Vision (ECCV)*, Amsterdam, The Netherlands, 2016.

[50] M. Devanne, H. Wannous , S. Berretti, P. Pala , M. Daoudi and A. D. Bimbo, "3-D Human Action Recognition by Shape Analysis of Motion Trajectories on Riemannian Manifold," *IEEE Transactions on Cybernetics*, vol. 45, no. 7, pp. 1340-1352, 2015.

[51] C. Wang , Y. Wang and A. L. Yuille, "Mining 3D Key-Pose-Motifs for Action Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016.

[52] C. Chen , R. Jafari and N. Kehtarnavaz, "Action recognition from depth sequences using depth motion maps-based local binary patterns," in *Winter Conference on Applications of Computer Vision (WACV)*, Hawaii, 2015.

[53] K. Jin , J. Min , J. Kong , H. Huo and X. Wang , "Action Recognition Using Vague Division Depth Motion Maps," *The Journal of Engineering*, vol. 2017, no. 4, pp. 77-84, 2017.

[54] G. Li , K. Liu , W. Ding , F. Cheng and B. Chen, “Key-Skeleton-Pattern Mining on 3D Skeletons Represented by Lie Group for Action Recognition,” *Mathematical Problems in Engineering,* vol. 2018, 2018.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Stacking Seismic Data Based on Ramanujan Sums

Karthikeyan Elumalai, *Member, IEEE*, Devendra Kumar Yadav, *Member, IEEE*, Anup Kumar Manpura, *Student Member, IEEE*, and R. K. Patney, *Senior Member, IEEE*

***Abstract*— Real seismic data consist of multiple traces captured by an array of receivers. These multiple traces are sorted by the common midpoint between the source and the receiver, and the time-lag between different traces is corrected by a normal move-out correction process. After these preprocessing steps, the sorted traces contain the same information about the earth sublayers and are corrupted by noise. The next step, termed stacking in seismic signal processing, involves the construction of an optimum trace with an improved signal-to-noise ratio (SNR) from these sorted traces. In this letter, we present an improved method for weighted stacking, where each trace is weighed in accordance with the noise variance. Using the first-order derivative property of Ramanujan sums, we perform the estimation of noise variance in each trace. Numerical results demonstrate that the method presented in this letter has better SNR for the trace obtained after stacking in comparison with existing methods.**

***Index Terms*— Denoising, Ramanujan's sums, stacking.**

I. Introduction

SEISMIC reflection is a cost-effective method of exploration to estimate the earth's subsurface layers that are rich in natural resources such as oil, minerals, water and so on [1]. In this method, a short seismic wavelet is transmitted from the surface of the earth. The seismic wavelet propagates downward into the different layers of the earth. Due to impedance contrast, the seismic wavelet is reflected toward the surface and these reflected waves are recorded by an array of geophones arranged on the earth surface. The geophone converts the wave motion into an electrical signal, commonly known as a seismic trace. The observed seismic trace contains information about seismic source, the reflection coefficients, and is corrupted by noise [2]. The experiment is repeated with small increments in the position of the seismic source and geophones at different time intervals. The geophone outputs are recorded for these repeated excitations, and the collection of these trace is called seismic data. There are various processing steps involved in obtaining the information of the sublayers, as shown in Fig. 1. In the preprocessing step, the amplitude loss in the trace due to layer absorption and geometric divergence is corrected [3]. In the acquired data, some of the geophones observe a wave that is reflected from the same point during different excitations by the source wavelet. Those geophone output traces are then sorted out based on the common midpoint (CMP) between the source and the receiver. The time delays between the sorted traces are also corrected by a normal move-out (NMO) method [4]. The set of NMO-corrected traces must be combined, as it contains the same information about the earth layers. Stacking in seismic processing is referred to the process of combining a collection of noisy NMO-corrected traces into a single trace. The process of stacking attenuates the noise and simultaneously amplifies the desired signal, leading to an overall improvement in the signal-to-noise ratio (SNR). As a result, stacking produces an optimum trace due to its effectiveness in noise and multiplies suppressing in order to get a coarse imaging that approximates the underground structure and provide foundational information for subsequent inversion, migration, and image.

In literature, there are various methods proposed to perform stacking on NMO-corrected seismic data to improve the SNR of the seismic trace. Conventional stacking [5], which is performed by averaging an NMO-corrected data set, is optimal only when the noise components in all traces are uncorrelated, are normally distributed, are stationary, and are of equal magnitude. Robinson [6] proposed a weighted stacking method to reduce noise. In this method, the optimum trace is computed as a linear combination of the NMO-corrected traces, and each trace is scaled by a weighting factor proportional to the signal-to-noise ratio (SNR) of the trace. Schoenberger [7] propose an optimum weighted stack for multiple suppressions, with weight determined by solving a set of optimization equations. Neelamani *et al.* [8] propose a stack-and-denoise (SAD) method, which exploits the structure of the seismic signals to obtain an enhanced stack. Zhang and Xu [9] present a high-order correlative weighted stacking technique on the basis of wavelet transformation and high-order statistics. Rashed [10] proposes smart stacking, which is based on optimizing the seismic amplitudes of the stacked signal by excluding harmful samples from the stack and applying larger weight to the central part of the sample population. In [11], time-dependent smooth weights are applied (which are taken as local correlation coefficients between the reference traces and the prestack traces). In all the aforementioned works, the reference trace (assign an equal weight for all the traces and combine) plays an important role, which is assumed less noisy. However, in practice, the reference trace may also contain noise so that the stacked traces contain noise. Limitations to the SNR ratio are common to all signal-extraction processes, and various

Manuscript received May 11, 2019; revised September 23, 2019; accepted October 23, 2019. *(Corresponding author: Karthikeyan Elumalai.)*

K. Elumalai is with the Department of Electronics and Communication Engineering, SRM University-AP, Amaravati 522502, India (e-mail: imkarthi@gmail.com).

D. K. Yadav is with DRDO, SAG, New Delhi 110054, India (e-mail: devhere@gmail.com).

A. K. Manpura is with the Department of Electrical Engineering, Delhi Technological University, New Delhi 110042, India (e-mail: anupiitd@gmail.com).

R. K. Patney is with the Department of Electrical Engineering, IIT Delhi, New Delhi 110016, India (e-mail: rkpatney@ee.iitd.ernet.in).

Digital Object Identifier 10.1109/LGRS.2019.2951300

1545-598X © 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



Fig. 1. Block diagram of seismic data processing.

methods are employed to bring about improvements. In order to overcome those problems, we have proposed a new method to estimate the noise variance from each trace based on Ramanujan's sum.

The rest of this letter is organized as follows. In Section II-A, we discuss the seismic signal model. In Section II-B, we describe the preliminaries of Ramanujan's sum. In Section III, we describe the stacking. In Section IV, we discuss about noise variance estimation, and in Section V, the accuracy of the proposed method is verified by the synthetic and real seismic data.

II. Mathematical Preliminaries

A. Signal Model

In this section, we present a mathematical model for the seismic trace [3]. Given a set of M seismic traces $y_1, y_2, \ldots, y_M$ from a CMP gather that have been NMO-corrected. Each trace $y_i \in \mathbb{R}^{N\times 1}$ is of length N. The traces can be arranged in an $N \times M$ data matrix Y as follows:

$$Y = \begin{bmatrix} y_1 & y_2 & \ldots & y_M \end{bmatrix}. \tag{1}$$

Each trace y_i is contaminated with additive Gaussian noise [11]. Therefore, Y can be written as a sum of two matrices: the first matrix containing the noise-free seismic trace and the second matrix consisting of noise. Let the noise-free matrix be $A = \begin{bmatrix} a_1 & a_2 & \ldots & a_M \end{bmatrix}$ where $a_i \in \mathbb{R}^{N\times 1}$ is the ith noise-free trace and $G \in \mathbb{R}^{N\times M}$ be the noise matrix. Therefore, we have

$$Y = A + G. \tag{2}$$

Each noise-free seismic trace a_i can be mathematically modeled as the convolution sum of reflection coefficients $r(n)$ and the seismic wavelet $w(n)$

$$a_i(n) = r(n) * w(n) \tag{3}$$

where $*$ represents the convolution operator and n represents the discrete time index.

B. Ramanujan Sum

The Ramanujan's sums, first introduced in [14], are trigonometric sums. Utilizing these trigonometric sums, expressions for a wide variety of algebraic functions can be obtained. For any positive integer q, the Ramanujan sum is defined as

$$c_q(n) = \sum_{\substack{k=1 \\ (k,q)=1}}^{q} \cos\frac{2\pi kn}{q}, \qquad \forall n \in \mathbb{Z}. \tag{4}$$

Here, $(k, q) = 1$ denotes the greatest common divisor (gcd) of k and q is 1. In other words, k and q are coprime. For example, if $q = 10$, integers that are less than and are coprime to q are $k = 1, 3, 7$, and 9. Therefore, $c_{10}(n)$ can be written as

$$c_{10}(n) = e^{j2\pi n/10} + e^{j6\pi n/10} + e^{j14\pi n/10} + e^{j18\pi n/10}. \tag{5}$$

Several standard arithmetic functions defined on a set of nonzero positive integers in the number theory can be expressed as a linear combination of Ramanujan sum, $c_q(n)$ []. Let $X(n)$ be any function of n, then we have

$$X(n) = \sum_{q=1}^{\infty} \alpha_q c_q(n), \quad n \geq 1 \tag{6}$$

where α_q are the Ramanujan sum's coefficients. It can be clearly seen that the Ramanujan sum $c_q(n)$ defined in (4) is q-periodic. A few of $c_q(n)$, for one period, are as follows:

$$\begin{aligned} c_1(n) &= 1 \\ c_2(n) &= 1, -1 \\ c_3(n) &= 2, -1, -1 \\ c_4(n) &= 2, 0, -2, 0. \end{aligned}$$

From the above equation, it is clear that the terms in $c_q(n)$ are always integer.

For a given q,($q \in 1$), we define the Ramanujan operator $\hat{R}_q$ as a linear, shift-invariant operator whose kernel is defined as one period of $c_q(n)$. We denote the kernel of this operator as $\hat{c}_q(n)$. Mathematically

$$\hat{c}_q(n) = \begin{cases} c_q(n), & 0 \leq n \leq q-1 \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

and the operator is defined as

$$(\hat{R}_q x)(n) = \sum_k \hat{c}_{d_j}(k) x(n-k).$$

In the next section, we apply this Ramanujan sum operator to estimate the noise variance form seismic trace.

III. Stacking

Stacking is performed on the CMP-gathered and NMO-corrected seismic traces to suppress the noise and enhance the signal components in the trace by combining all the traces into a single trace [11], [17]. Mathematically, the process of equal weight stacking can be expressed as follows:

$$ST = (1/k)\sum_{i=1}^{M} y_i \tag{8}$$

where M is the number of seismic traces and ST is the averaged trace obtained by the stacking process. However, the quality of each trace is different and each trace should contribute differently to the final stacked trace. A better stacking approach is to weigh each trace before stacking based on certain criteria. The weighted stacking process introduced in [6] can be mathematically expressed as

$$\text{STW} = \frac{1}{\sum_i w_i}\sum_{i=1}^{M} w_i y_i \tag{9}$$

where $w_i(t)$ is the weight applied to trace i and time t in a CMP gather, which is determined by noise variances $w_i = 1/\sigma_i$. The STW is the stacked trace after weighting.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

IV. Noise Variance Estimation

In the previous section, the signal model was discussed, where each trace is contaminated with the Gaussian noise. In this section, we will elaborate the method for the estimation of the noise variance in each trace. The basic assumption here is that noise lies in the detailed part of the signal. Yadav *et al.* [15] have demonstrated that Ramanujan sum operating on a function acts as a derivative. Since derivatives are sensitive to noise, the Ramanujan operator can be used to estimate the noise level. The method for the estimation of noise variance using the Ramanujan operator is also discussed in [15].

In the previous section, the seismic signal model and the weighted stacking are discussed. In order to perform weighted stacking, the weights need to be known. For the CMP-gathered and NMO-corrected seismic traces, the noise variance can be estimated based on the Ramanujan sum. This can be achieved by considering the kernel of the corresponding Ramanujan operators, each corresponding to the divisor of the length of the input signal. Let $c_{\hat{d}_j}$ be the kernel of the Ramanujan operator corresponding to the divisor d_j, for $j = 1, \ldots, k$. The application of this operator on a seismic trace y will result in the derivative of the trace

$$(\hat{R_{d_j}} y)(n) = \sum_k \hat{c}_{d_j}(k) y(n-k). \tag{10}$$

Downsampling $(\hat{R_{d_j}} y)(n)$ by a factor of d_j will represent the detailed component of that seismic trace. Mathematically

$$y_{d_j}(n) = (\hat{R_{d_j}} y)(n d_j).$$

This computation needs to be performed for each divisor and the detailed component for each divisor can then be calculated similarly. Let there be k divisors, then, the noise variance can be estimated as [16]

$$\hat{\sigma} = \sum_{j=1}^{k} \frac{\mathrm{MAD}(y_{d_j})}{0.6745k} \tag{11}$$

where MAD(.) is the mean absolute deviation [15] and $\hat{\sigma}$ denotes the estimated noise variance.

V. Results and Discussion

In this section, we study two cases to demonstrate the effectiveness of the proposed method. The proposed method was applied on both CMP-gathered and NMO-corrected synthetic and field data. The synthetic seismic data are generated using (3). The stacking performed on the seismic data was carried out using (9). In order to perform weighted stacking, the weights were estimated for each trace using (11). To compute the performance using the proposed approach numerically, we use the SNR defined as [11]

$$\mathrm{SNR} = 10\log_{10}\frac{\sum_t T^2(t)}{\sum_t [NT(t) - T(t)]^2} \tag{12}$$

where $T(t)$ is the noise-free trace and $NT(t)$ is the noisy trace in the CMP gather. The mean squared error (MSE) of the weighted stacking trace $\hat{ST}(i)$ and the noiseless trace $LT(i)$ for $i = 0, \ldots, q$ is given as

$$\mathrm{MSE} = \frac{1}{q}\sum_{i=0}^{q}\left(LT(i) - \hat{ST}(i)\right)^2.$$

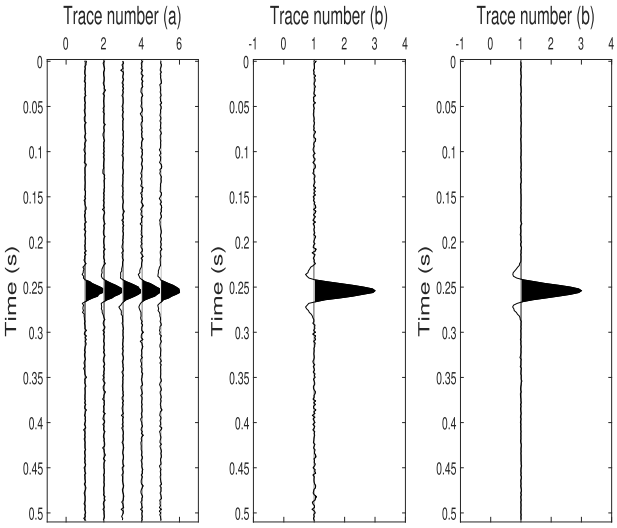


Fig. 2. Weighted stacking on synthetic seismic data.

TABLE I
Comparison of Various Weighted Stacking Methods of SNR and MSE

Methods	SNR (dB)	MSE
Conventional method	8.4 dB	0.091
Local correlation method	13.5 dB	0.027
Proposed method	15.2 dB	0.001

In next two sections, we study the performance of the proposed stacking method on the synthetic data and the real field data. The results are then compared with existing methods in literature [11] on stacking.

A. Application of Proposed Stacking Algorithm on Synthetic Data

In this letter, we consider synthetic seismic data to perform stacking. The data consist of five traces $(y_1, y_2, \ldots, y_5)$, and each trace has 200 time-sample points with a sampling interval of 2 ms. The traces are shown in Fig. 2(a). The white Gaussian random noise was added with all the five traces. All the five traces are combined based on the conventional method and obtained a single trace shown in Fig. 2(b). The proposed stacking method is applied to the NMO-corrected trace to suppress noise, which is present in each trace and enhances the signal component, as shown in Fig. 2(c). The evaluation methodology in this setting is the same as the previous case study [11]. The proposed stacking method compared with the existing method [11] is shown in Table I. From the table, the proposed stacking method suppresses the noise effectively and the SNR of the seismic trace improved while compared with the other method.

B. Application of Proposed Stacking Algorithm on Real Field Data

In this section, we demonstrate the proposed stacking method on the real data set. The data for the experiments have been taken from [18]. The raw seismic data follow the processing steps as given in the block diagram 1. The reflected

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

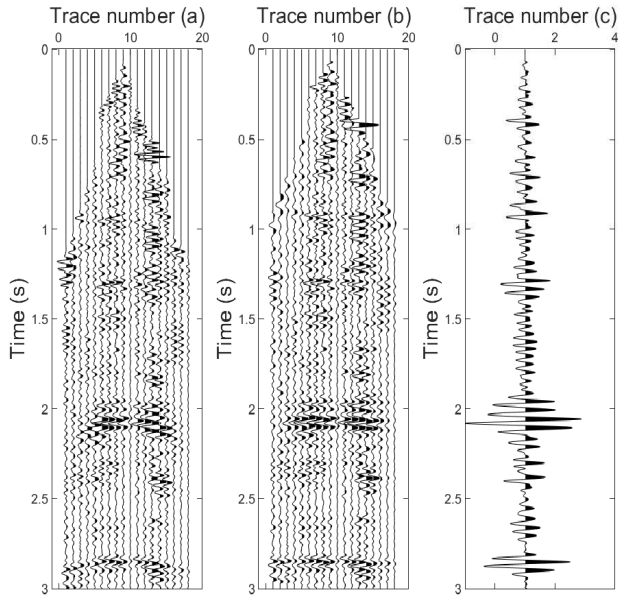


Fig. 3. Weighted stacking on real seismic data.

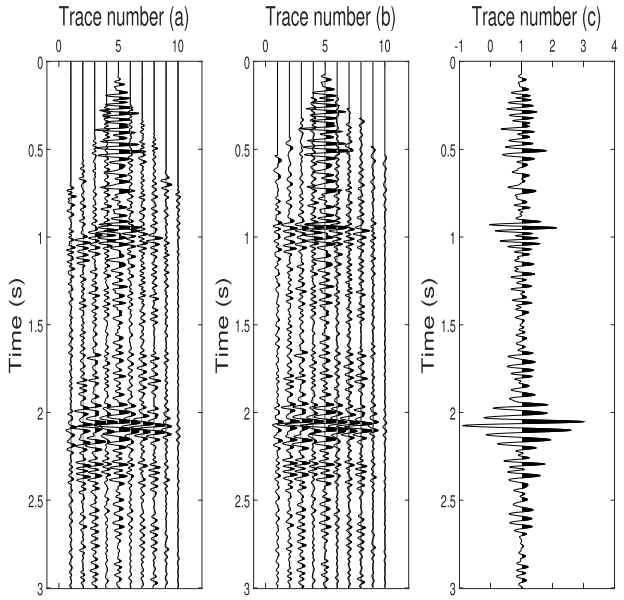


Fig. 4. (a) Before NMO correction. (b) After NMO correction. (c) Traces are weighted stacked using proposed method.

seismic wave is recorded for 3 s. The real data were segmented based on CMP gather.

Case 1: In this case, we consider the ℓth CMP gather containing only 17 traces, and each contains 1500 time-sample points. The sampling rate is 500 Hz. The traces in the ℓth CMP gather are misaligned, as shown in Fig. 3(a). The time difference corrected by NMO correction is shown in Fig. 3(b). The result of the proposed stacking method is applied to the ℓth CMP-gathered and NMO-corrected data. The weight was estimated from each trace separately using (11), and the traces are stacked using (9). As can be observed, the large amplitude reflections in the data correspond to the reflections at 2.15 and 2.8 s, while that for small reflections are visible at 1.25 and 2.25 s, as shown in Fig. 3(c). The large reflection represents hard rock in the subsurface.

Case 2: In this case, we considered the kth CMP gather from the same real data. In this gather, there are ten traces, as shown in Fig. 4. The traces are combined by using (9). The weights are estimated based on the proposed method using (11). The small reflections are clearer in the stacked trace at 0.95 and 2.1 s, as shown in Fig. 4(c).

VI. Conclusion

The weighted stacking approach can obtain a much improved stacking result compared with the equal-weight stacking, considering the increased SNR. We proposed a new approach for the weighted stacking method that is based on Ramanujan's sum. The proposed approach is tested by a synthetic CMP gather and a field data example, which shows very promising performance.

Acknowledgment

The authors would like to thank Prof. S. D. Joshi, who helped us to know about Ramanujan sums.

References

[1] G. Nolet, *Seismic Tomography: With Applications in Global Seismology and Exploration Geophysics*, vol. 5. Springer, 2012.
[2] D. W. Steeples and R. D. Miller, "Seismic reflection methods applied to engineering, environmental, and ground-water problems," in *Proc. Geotech. Environ. Geophys.*, vol. 1, 1988, pp. 409–461.
[3] Ö. Yilmaz, *Seismic Data Analysis*, vol. 1. Tulsa, OK, USA: Society of Exploration Geophysicists, 2001.
[4] P. Deng, Y. Chen, Y. Zhang, and H.-W. Zhou, "Weighted stacking of seismic AVO data using hybrid AB semblance and local similarity," *J. Geophys. Eng.*, vol. 13, no. 2, pp. 152–163, 2016.
[5] W. H. Mayne, "Common reflection point horizontal data stacking techniques," *Geophysics*, vol. 27, no. 6, pp. 927–938, 1962.
[6] J. C. Robinson, "Statistically optimal stacking of seismic data," *Geophysics*, vol. 35, no. 3, pp. 436–446, 1970.
[7] M. Schoenberger, "Optimum weighted stack for multiple suppression," *Geophysics*, vol. 61, no. 3, pp. 891–901, 1996.
[8] R. Neelamani, T. A. Dickens, and M. Deffenbaugh, "Stack-and-Denoise: A new method to stack seismic datasets," in *Proc. 76th SEG Annu. Int. Meeting*, 2006, pp. 2827–2831.
[9] S. Z. Zhang and Y. X. Xu, "Higher-order correlative stacking for seismic data in the wavelet domain," (in Chinese), *Chin. J. Geophys.*, vol. 49, no. 2, pp. 554–560, 2006.
[10] M. A. Rashed, "Smart stacking: A new CMP stacking technique for seismic data," *Lead. Edge*, vol. 27, no. 4, pp. 462–467, 2008.
[11] G. Liu, S. Fomel, L. Jin, and X. Chen, "Stacking seismic data using local correlation," *Geophysics*, vol. 74, no. 3, pp. V43–V48, 2009.
[12] Q. Li and J. Gao, "Application of seismic data stacking in time–frequency domain," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 9, pp. 1484–1488, Sep. 2014.
[13] D. G. Stone, "Wavelet estimation," *Proc. IEEE*, vol. 72, no. 10, pp. 1394–1402, Oct. 1984.
[14] S. Ramanujan, "On certain trigonometrical sums and their applications in the theory of numbers," *Trans. Cambridge Philos. Soc.*, vol. 22, no. 13, pp. 259–276, 1918.
[15] D. K. Yadav, G. Kuldeep, and S. D. Joshi, "Ramanujan sums as derivatives and applications," *IEEE Signal Process. Lett.*, vol. 25, no. 3, pp. 413–416, Mar. 2018.
[16] F. R. Hampel, "The influence curve and its role in robust estimation," *J. Amer. Statist. Assoc.*, vol. 69, no. 346, pp. 383–393, 1974.
[17] J. Xie, W. Chen, D. Zhang, S. Zu, and Y. Chen, "Application of principal component analysis in weighted stacking of seismic data," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 8, pp. 1213–1217, Aug. 2017.
[18] W. A. Mousa and A. A. Al-Shuhail, *Processing of Seismic Reflection Data Using MATLAB* (Synthesis Lectures on Signal Processing), vol. 5, no. 1. 2011, pp. 1–97.

© The British Computer Society 2019. All rights reserved.
For permissions, please email: journals.permissions@oup.com
doi: 10.1093/comjnl/bxz103

Stochastic Gradient Descent–Whale Optimization Algorithm-Based Deep Convolutional Neural Network To Crowd Emotion Understanding

Avinash Ratre

Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, 110042, India

ratreavinash9@gmail.com; ratreavinash@gmail.com

Crowd emotion understanding is an interesting research area that assists the security personnel to read the emotion/activity of the crowd in the locality. Most of the traditional methods utilize the low-level visual features to understand the crowd emotions that extend the gap between the low- and the high-level features. With the aim to develop an automatic method for emotion recognition, this paper utilizes the deep convolutional neural network (deep CNN). For the effective emotion recognition, it is essential to select the key frames of the video using the wavelet-based Bhattacharya distance. The key frames are fed to the space-time interest points descriptor that extracts the features and forms the input vector to the classifier. Deep CNN is trained using the proposed Stochastic Gradient Descent–Whale Optimization Algorithm, which is the unification of the standard stochastic gradient descent algorithm with whale optimization algorithm. The proposed classifier recognizes the emotions of the crowd, such as angry, escape, fight, happy, normal, running/walking and violence. The analysis of the method proved that the proposed approaches acquired a maximal accuracy, specificity and sensitivity of 0.9693, 0.9936 and 0.9675, respectively.

Keywords: optimization; deep learning; crowd behavior understanding; surveillance videos; crowd emotion recognition

Received 21 November 2018; Revised 23 March 2019; Editorial Decision 15 May 2019
Handling editor: Yannis Manolopoulos

1. INTRODUCTION

1.1. Background of video surveillance

Video surveillance attracts the researchers as security is the major concern to the community and performing the crowd analysis ensures the automatic analysis and detection of the normal/abnormal events occurring in day-to-day life of the public. The crowd video analysis initially identifies the individual and tracks them with respect to time [1]. Particularly, the security concerns of the society become effective through the crowd emotion recognition (CER) mainly, in the crowded scenes, like public meetings, competition spots, demonstrations and so on [2, 3]. Emotions are the feelings that induce stimuli around the environment or sometimes provoke the internal body functions that modulate the behaviors as actions. Emotion is a Latin word meaning 'move' and the emotions are based on the situation of the environment [4]. Recognizing the emotions using the computer is applicable in various marketing and advertisements related to the specific user profiles. For instance, recognizing the emotions of the supporters from the large crowd in the stadium is a huge task to identify the actual emotion of the supporters regarding the win and loss of the team, injuries of their favorite player, when their hero hits an excellent shot and so on [5]. Thus, performing the emotion recognition from the crowded videos plays a vital role in examining the individual enactments, security, during the impact analysis of an event and so on [4].

1.2. Need for CER

CER is progressed as two major steps; in the first step, the local features of the crowd is obtained, which is then subjected

Downloaded from https://academic.oup.com/comjnl/advance-article-abstract/doi/10.1093/comjnl/bxz103/5627747 by Delhi Technological University user on 05 December 2019

to the detection of the casual relationship in the second step. Thus, during the identification of the casual relationships, the emotions of the crowd are classified using any classifier [6]. The relationships are established based on the happenings in the environment and the interactions between them [7]. In the field of psychology, emotions of persons are referred and the basic emotions or classes include the happy, angry, sad, fear, disgust, surprise and so on [8]. Emotion recognition, in general, is based on the databases used, which focuses on different types of the emotions. Thus, the emotion recognition methods are broadly grouped depending on the total number of subjects in the sample as static- and dynamic-based methods [9].

1.3. Methods used for CER

There are a lot of supervised learning techniques to perform the emotion recognition and some of them include neural network (NN) [10], deep belief networks (DBNs) and hidden Markov model (HMM) [11], which are engaged in recognition at various environmental settings including the closed, indoor and outdoor settings [12]. Researchers on the other hand used the Bayesian networks for monitoring the crowd [13, 14]. Additionally, models, like support vector machine (SVM) [15], Gaussian mixture model (GMM) and recurrent NNs, are used for emotion recognition. There are also other classifiers in the literature, like brain emotional learning model (BEL) that used the adaptive neuro-fuzzy inference system and multilayer perceptron to recognize the emotion. Yet another method for recognizing the emotions is a multiple kernel Gaussian process classification [16].

1.4. Overview of the proposed recognition approach

Learning-based methods for CER gain advantages among the researchers. Accordingly, this work aims to develop a recognition approach based on crowd emotion understanding using deep learning. The complete process is done using three components. In the first component, the important frames are identified from the video using wavelet-based Bhattacharyya distance matching. Once the important frames are identified, space-time interest (STIs) points are identified as the feature vector. Then, the recognition is done using deep convolutional neural network (deep CNN), which is trained based on a hybrid training algorithm. The hybrid training algorithm is developed by combining the stochastic gradient descent (SGD) with whale optimization algorithm (WOA) [26, 27] for the optimal selection of weights in the deep CNN. Thus, the proposed approach with the hybrid optimization-based deep CNN performs the CER.

Contribution of the research. The contributions of the paper are as follows:

- ***Extraction of the interesting points for accurate classification:*** The application of the STIs descriptor to the key frames for obtaining the interest points in order to facilitate the accurate emotional recognition using the crowded videos.
- ***Optimization-enabled deep learning NN for recognition:*** The CER using the deep CNN classifier is the contribution that is tuned optimally using the proposed SGD–WOA, which is the modification of the WOA with the SGD update equation. The advantage is that the convergence of the developed algorithm is high when compared with the other optimizations.

The structure of the paper is as follows: review of the literatures in part 2, proposed method of recognition in part 3 and part 4 evaluates the results. Lastly, part 5 summarizes the presented work.

2. LITERATURE REVIEW

In this section, the review of the eight iterature works are presented. In 2014, Baig *et al*. [4] developed the emotion detection using bio-inspired model that used the dynamic Bayesian networks to predict the emotions of the crowd. The system was effective in predicting the crowd situation prior to happening. The drawback of the method was that the accuracy dropped when there was a problem associated with classification, mainly due to the classification of immobile crowds. In 2016, Rabiee *et al*. [17] developed an approach based on the attributes to describe the emotions of the crowd that performed the classification with better discrimination. The demerit of the method was regarding the training samples and it failed for large number of crowd datasets. In 2016, Patwardhan [14] developed an edge-based grid super-imposition method that extracted the feature. The method offered better performance and did not suffer from training data insufficiency. However, the method was not effective for the outdoor bigger crowd. In 2014, Baig *et al*. [18] developed a dynamic probabilistic model to extract the dynamic patterns from the events in the video and the patterns described the behavior of the crowds. The method was capable of determining the emotions of persons from individual to the collective degree with a better accuracy, but the method failed due to the collective datasets. In 2015, Baig *et al*. [6] employed the probabilistic models for recognizing the emotions that predicted the emotion of the person from the surveillance areas. The dynamic pattern of the persons varied based on their relationship in the environment and their past interactions. Moreover, the method distinguished between the negative and positive emotions, but was inapplicable for different situations. In 2016, Urizar *et al*. [19] used the hierarchical Bayesian network to describe the human behaviors, which rendered effective results, but failed to recognize the presence of interactions. In 2016, He and Zhang [20] used convolutional NN for emotion recognition that was capable of rendering the active level features with higher recognition accuracy. However, the use of optimization strategies could

Downloaded from https://academic.oup.com/comjnl/advance-article-abstract/doi/10.1093/comjnl/bxz103/5627747 by Delhi Technological University user on 05 December 2019

further refine the performance of the approach. In 2018, Liu *et al*. [28] developed a speech emotion recognition technique that depends on the refined BEL framework. BEL was inspired by the emotional processing mechanism of the limbic system in the brain. Here, genetic algorithm was used to update the weights of BEL framework. When compared with the conventional speech emotion recognition, this method had high performance. In 2018, Yan *et al*. [29] introduced a multi-cue fusion emotion recognition framework by replicating human emotions from three interrelated cues, such as audio signal, facial landmark activity and facial texture and then merging them together. Here, a four convolutional layer network was intended for learning the emotion related patterns. The performance of this method was analyzed on two challenging databases, and the results showed that this method was competitive and valuable.

3. CER USING THE DEEP LEARNING CLASSIFIER WITH THE STI POINTS FEATURES

Problem statement: In the CER, the main challenges lie in capturing of emotions from the video camera [4], filling the space between the low-level features obtained from images and concepts regarding the high-level semantics [20]. There is a requirement of a large number of training images and the recognition is effective in the presence of the static or the dynamic image in the video that vary between the frames. Additionally, it is ineffective for the real-time applications, where there is dominancy in the dynamic objects [21]. Most of the current recognition systems engaged in recognizing the emotions are inapplicable in the realistic applications, due to the fact of using the physiological signal for emotion recognition. The traditional methods failed to map the various physiological patterns lying in the different types of the emotions and in addition, the physiological data are highly sensitive to the motion artifacts [22]. Even though the emotion recognition using the HMM [13] gains a lot of interest, it lacked feasibility and was tough to implement due to the requirement of highly intense model for training and evaluating the dependency among the large mass of the training data. Moreover, the need for the emotion recognition is to predict the emotions of the crowd and ensure some precautions about what is going to happen in future. The main drawback associated with detecting the emotions of crowd is that since we use the video cameras, the features, like the face emotion activity and voice features, become ineffective. As those features suffer from the mixing up of voices, lack of the history of the crowd, confusion regarding the posture and movement in crowd and so on, there is a need for the effective emotion recognition method. With this in mind, the automatic CER is progressed using the deep CNN classifier. Thus, in the proposed method, we have improved the accuracy, sensitivity and specificity of the system by implementing SGD–WOA algorithm.

3.1. Methodology

Initially, the key frames in the video are chosen using the Bhattacharya distance between the frames, for which initially the wavelet transform (WT) of the individual frames is determined. The low-frequency band of the WT is subjected to the key frame selection using the minimum value of the Bhattacharya distance. The key points are obtained from the selected key frames using the STI descriptor and the image with the key points is given as input to the deep CNN classifier, which is tuned optimally using the proposed SGD–WOA. Figure 1 shows the architecture of proposed CER using the deep CNN classifier. Consider the video V with n number of frames, given as

$$V = \{f_1, f_2, \ldots, f_i, \ldots, f_n\}, \tag{1}$$

where f_i indicates the i^{th} frame in the video V.

3.2. Application of WT to the individual frames

The individual frames in the video are subjected to the WT so that the frequency bands are extracted from the individual frames. The WT of the frames yields the highly precise information of the data such that it decomposes the signal as the function of time.

$$\left[f_i^{LL}, f_i^{LH}, f_i^{HL}, f_i^{HH}\right] = D\ [f_i], \tag{2}$$

where $f_i^{LL}, f_i^{LH}, f_i^{HL}, f_i^{HH}$ are the low- and high-frequency bands of the i^{th} frame of the video V. The low-frequency band contains the significant information of the image and so it is chosen for further processing. The LL-band of all the frames is chosen and histogram of the band for frames is determined. The size of the individual histogram is $[1 \times 256]$.

The histogram of the wavelet corresponding to i^{th} frame is denoted as H_i^{LL}. The histogram of the LL bands of all the frames is

$$H = \{H_1, H_2, \ldots., H_i, \ldots, H_n\}. \tag{3}$$

These frames are fed to the key frame selection step that is progressed using the Bhattacharya coefficient (BC).

3.3. BC for the selection of the key frames

The significance of the key frame selection is to minimize the complexity and to find the highly significant frame that yields valuable information to perform the classification. The key frame selection criterion facilitates the detection of the frame motion parameters, effects, audio and other properties that vary with time. Hence, the key frame maps an object with respect to time, where the property attributes, such as spatial position, audio volume or opacity, is specified, followed with the interpolation of the values among the key frames. The

Downloaded from https://academic.oup.com/comjnl/advance-article-abstract/doi/10.1093/comjnl/bxz103/5627747 by Delhi Technological University user on 05 December 2019

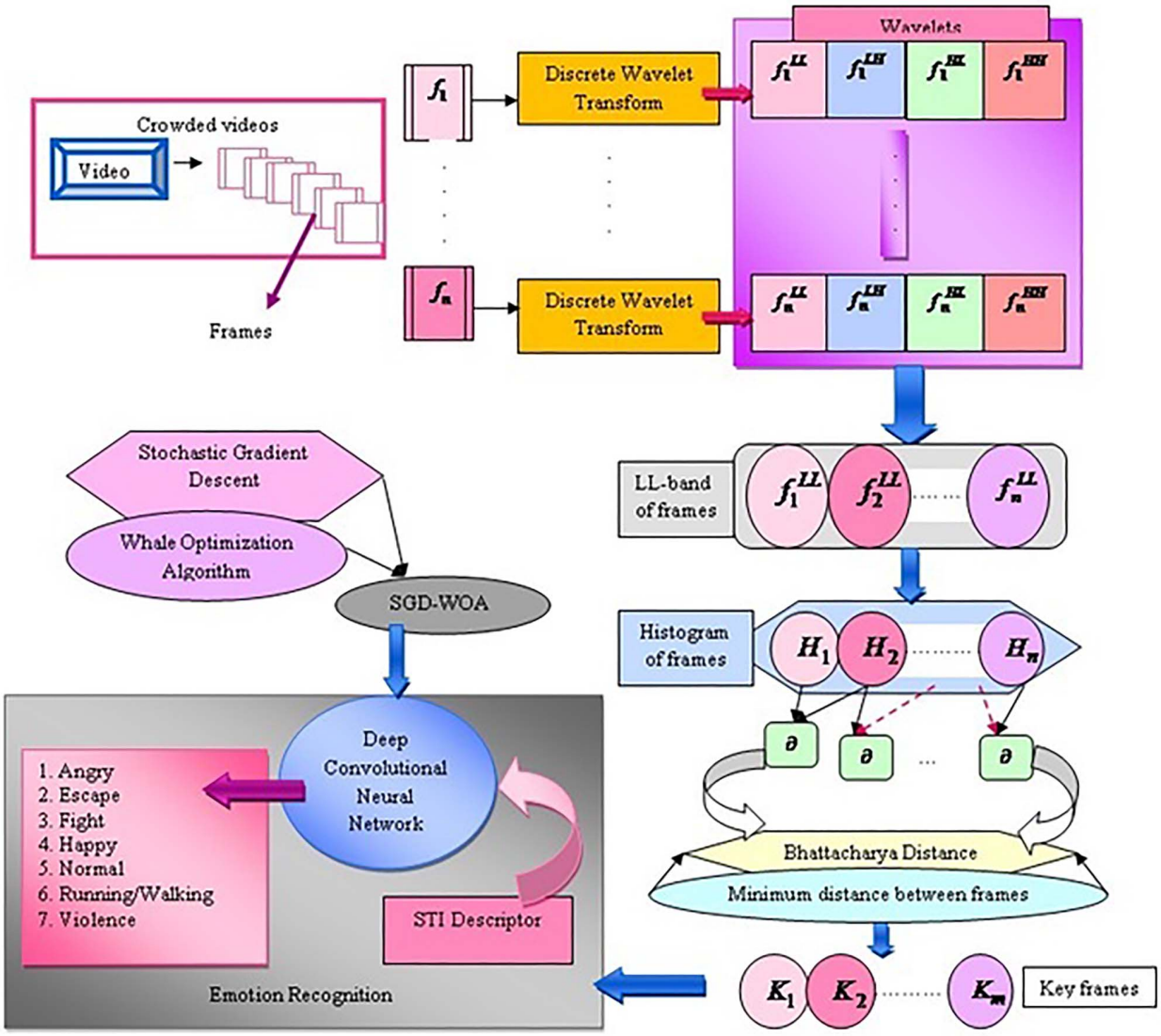


FIGURE 1. Schematic diagram of emotion recognition.

key frames indicate the change over time, and these changes describe the position/state of the object. The distance between two key frames is computed to decide the key frame that describes the state of object in the beginning and the new state of the object at the end or termination of the movement of the object. The BC computes the similarity among two frames describing the existing overlap among the population. The BC uses the relative closeness between two frames and this measure is highly reliable than any other similarity measures. The similarity measure is determined based on the formula shown in equation (4) that determines the similarity between the frame H_i and the neighboring frame H_{i+1}.

$$\partial\,(H_i, H_{i+1}) = \sum_{i=1}^{n} \sqrt{(H_i)\,(H_{i+1})}, \tag{4}$$

where $\partial(H_i, H_{i+1})$ refers to the BC of the two frames. The selected number of the key frames determined using the BC is given as

$$K = \{K_1, K_2, \ldots, K_i, \ldots, K_m\}\,; \;\; m < n, \tag{5}$$

Downloaded from https://academic.oup.com/comjnl/advance-article-abstract/doi/10.1093/comjnl/bxz103/5627747 by Delhi Technological University user on 05 December 2019

where m denotes the number of the selected frames from the video.

3.4. Representation of the interesting points in the frame using the STI descriptor

The STI features [23] ensure the extraction of the motion patterns belonging to the primitive events that result due to the movement of the object, which is demonstrated between the frames. The local features are determined from the frame $K_i(u, v, \delta)$ based on the small-scale representation. The features related to the scale and the velocity of the object are determined to compute the invariance of the object concerning the size of the object in the frame. The motion and the spatial appearance of the actions/activity in the frame are extracted from the local features of the spatio-temporal neighborhoods. To compute the local features, the spatial temporal jets are computed and are denoted as

$$J = (\mathrm{M}_u, \mathrm{M}_v, \mathrm{M}_\delta, .., \mathrm{M}_{\delta\delta\delta}) \,. \tag{6}$$

Thus, the STI features denote the interesting points in the frame and the feature vector is comprised of the interesting points in the image so that the emotion of the object in the frame is recognized effectively using the classifier. Let us represent the output from the STI descriptor as P.

3.5. Emotion recognition using the SGD–WOA-based deep CNN

The deep CNN classifier classifies the interesting points in the individual frames and derived the classes, which include angry, escape, fight, happy, normal, running/walking and violence. For the effective classification of the images, deep CNN classifier is used and the weights biases are determined using the proposed algorithm optimally.

3.5.1. *Structure of the deep CNN*

Deep CNN plays a major role in analyzing the images for better classification outcomes and it outperforms the other classifiers. The difference of deep CNN from the NN is that, in the NN, a single neuron is connected between the neurons in the successive layers, whereas in deep CNN, a patch of neuron in a layer is coupled with the individual neuron in the next layer. The architecture of deep CNN [24, 25] possess three layers, such as pooling (POOL) layers, convolutional (conv) layers and a Fully Connected (FC) layer as shown in Fig. 2. Each of the layers in the deep CNN perform specific functions including the development of the feature maps in the conv layers, sub-sampling the feature maps in the POOL layers and classification in the FC layer. With the usage of a large number of conv layers, the accuracy of the classification is enhanced.

Conv layers: The responsibility of the conv layers ($32 \times 32 \times 3$) relied on obtaining the patterns buried in the input feature vector using the conv filters that are connected using the

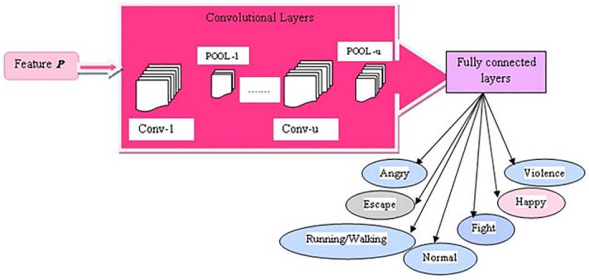


FIGURE 2. Architecture of deep CNN.

receptive fields, which act as an interconnection between the neurons in the previous layer with the successive conv layers through the trainable weights. The feature maps are developed through the convolution of the input feature vector with the trainable weights in such a way that the trainable weights are obtained using the proposed SGD–WOA approach. The neurons of the single layer engage themselves in extracting the variable features available in various location based on the variable weights of the single layer. Let us assume the input to the deep CNN is P and hence, the output from the conv layer is given as

$$\left(P_k^l\right)_{p,q} = \left(B_k^l\right)_{p,q} + \sum_{r=1}^{\kappa_1^{r-1}} \sum_{\tau=-\kappa_1^l}^{\kappa_1^l} \sum_{\upsilon=-\kappa_2^l}^{\kappa_2^l} \left(\varpi_{k,r}^l\right)_{\upsilon,\tau} * \left(P_r^{l-1}\right)_{p+\tau,q+\upsilon}, \tag{7}$$

where $*$ refers to the convolutional operator that paves way for obtaining the local patterns from the alternative conv layers and $\left(P_k^l\right)_{p,q}$ refers to the fixed feature map or the output from the l^{th} conv layers centered as (p, q). The output from the previous $(l-1)^{th}$ layer devises the input to the l^{th} conv layer. Let the weights of the conv layers be denoted as $\varpi_{k,r}^l$, which is the weights of l^{th} conv layer and the bias of l^{th} conv layer is denoted as B_k^l. Let us assume that there are u number of conv layers ($1 \leq l \leq u$) and the notations r,τ and υ denote the feature maps, which are the output from the individual conv filter obtained through scanning the input image pixel-wise, and the size of the conv filters is based on the number of pixels in the input image. The neurons in conv layers are organized in three-dimensions along the width, height and depth so as to extract the features from all the dimensions and the rectified linear unit (ReLU) layer uses an element-wise function to simplify the computation through the removal of the negative values. The output from the j^{th} ReLU layer is the activation function (Afn) of the $(l-1)^{th}$ previous layer and is given as

$$P_k^l = Afn\left(P_k^{l-1}\right). \tag{8}$$

Downloaded from https://academic.oup.com/comjnl/advance-article-abstract/doi/10.1093/comjnl/bxz103/5627747 by Delhi Technological University user on 05 December 2019

The importance of ReLU layer is regarding the speed of deep CNN, which is enhanced and offers the ability to deal with a large number of networks.

POOL layers: The POOL layer, i.e. a non-parametric layer, which is the layer with no bias and weights so that the main role of POOL is to undergo a fixed operation.

FC layers: The abstract features from the POOL are fed to the FC layer. At the termination of the network, the image is transformed as a single vector that signifies the classes of the image. The output of the FC layer is given as

$$S_k^l = \delta\left(P_k^l\right) \text{ with}$$
$$P_k^l = \sum_{r=1}^{\kappa_1^{r-1}} \sum_{\tau=-\kappa_1^l}^{\kappa_1^l} \sum_{\upsilon=-\kappa_2^l}^{\kappa_2^l} \left(\varpi_{k,r}^l\right)_{\upsilon,\tau} \left(P_r^{l-1}\right)_{p+\tau,q+\upsilon}. \quad (9)$$

3.5.2. SGD–WOA for optimal weight update in deep CNN

The optimal tuning of the weights in the deep CNN facilitates the better extraction of the feature maps and ensures the effective classification accuracy. The proposed SGD–WOA is the unification of the SGD [26] in the WOA [27] so that the developed algorithm offers better benefits to classification. Basically, WOA depends on the hunting strategy of humpback whales (H-whales), which is a population-based meta-heuristics optimization. The difference between WOA and the existing optimizations is that WOA determines the optimal search agent by chasing the prey and the presence of the spiral bubble net-attacking strategy. Thus, WOA is said to offer the convergence to the local optimum solution and is capable of dealing with the real solutions associated with the unknown search space [35]. The integration of SGD increases the efficiency that formulates the result based on the entire training set. However, the algorithm takes longer time to converge, which is nullified using WOA that renders greater convergence rate.

The SGD–WOA involves the encircling phase, in which the search agent tends to update the position based on the current best solution, which is described as

$$\overrightarrow{w_{z+1}} = \left|\overrightarrow{w_z^*} - \overrightarrow{L} \cdot \overrightarrow{G}\right|, \quad (15)$$

where $\overrightarrow{L}$ is the coefficient vector, z refers to the current iteration, $||$ defines the absolute value, $\overrightarrow{w_z^*}$ indicates the current best solution of the search agents at iteration z and $\overrightarrow{w_{z+1}}$ indicates the position of the search agent in the next iteration. $\overrightarrow{G}$ refers to the distance vector mentioning the distance between the search agent and the prey, which is updated based on the current position of the search agent, current best solution and the coefficient vector. Hence, the position update of the search agent in the encircling phase is given as

$$\overrightarrow{w_{z+1}} = \overrightarrow{w_z^*} - \overrightarrow{L} \cdot \left[\overrightarrow{E} \cdot \overrightarrow{w_z^*} - \overrightarrow{w_z}\right], \quad (11)$$

where $\overrightarrow{E}$ is the coefficient vector and on rearranging the above equation, we get

$$\overrightarrow{w_{z+1}} = \overrightarrow{w_z^*} - \overrightarrow{L} \cdot \overrightarrow{E} \cdot \overrightarrow{w_z^*} + \overrightarrow{L} \cdot \overrightarrow{w_z} \quad (12)$$

$$\overrightarrow{w_{z+1}} = \overrightarrow{w_z^*} \left[1 - \overrightarrow{L} \cdot \overrightarrow{E} \cdot\right] + \overrightarrow{L} \cdot \overrightarrow{w_z}. \quad (13)$$

The coefficient vectors are described based on the random vector and the optimization parameter $\overrightarrow{x}$ that decreases from 2 to 0 during the advent of exploration and exploitation. The random vector $\overrightarrow{e}$ varies between 0 and 1.

$$\overrightarrow{L} = 2\,\overrightarrow{e} \cdot \overrightarrow{x} - \overrightarrow{e} \quad (14)$$

$$\overrightarrow{E} = 2 \cdot \overrightarrow{x}. \quad (15)$$

$\overrightarrow{w_z^*}$ refers to the best solution, $\overrightarrow{w_z}$ refers to the position vector at time z and $\overrightarrow{x}$ refers to the random number. The exploitation phase commences with the decrease in the coefficient vector $\overrightarrow{L}$ varying in $[-e, e]$, where e decreases from 2 to 0 in every iteration. Therefore, the position of the search agent is found in a place among the current best solution and the original position of the search agent. During the exploration phase, the spiral shape movement of the H-whale is considered using the search agent or the H-whale that locates the position of the prey. Since the movement is described as the helix movement, the distance of the prey from the search agent is considered and the position update is given as

$$\overrightarrow{w_{z+1}} = \overrightarrow{E}.e^{da} \cdot \cos(2\pi a) + \overrightarrow{w_z^*}, \quad (16)$$

where $\overrightarrow{E} = |\overrightarrow{w_z^*} - \overrightarrow{w_z}|$ refers to the distance vector and after substituting the distance in the above equation, we get

$$\overrightarrow{w_{z+1}} = \left|\overrightarrow{w_z^*} - \overrightarrow{w_z}\right|.e^{da} \cdot \cos(2\pi a) + \overrightarrow{w_z^*}. \quad (17)$$

The modulus from the above equations is removed with the simple assumption $\overrightarrow{w_z^*} > \overrightarrow{w_z}$ so that the equation (17)

$$\overrightarrow{w_{z+1}} = \overrightarrow{w_z^*}e^{da} \cdot \cos(2\pi a) - \overrightarrow{w_z}.e^{da} \cdot \cos(2\pi a) + \overrightarrow{w_z^*}, \quad (18)$$

where $d$and$a$ indicate the shape of the spiral and random number in $[-1, 1]$. Similarly, $[\cdot]$denotes the element-wise multiplication. Thus, the position update of the H-whales is as

Downloaded from https://academic.oup.com/comjnl/advance-article-abstract/doi/10.1093/comjnl/bxz103/5627747 by Delhi Technological University user on 05 December 2019

follows:

$$\overrightarrow{w_{z+1}} = \begin{cases} \overrightarrow{w_z^*}\left[1 - \overrightarrow{L}\cdot\overrightarrow{E}\cdot\right] + \overrightarrow{L}\cdot\overrightarrow{w_z}\ ; & \rho < 0.5 \\ \overrightarrow{w_z^*}e^{da}\cdot\cos(2\pi a) - \overrightarrow{w_z}.e^{da} & \\ \quad\cdot\cos(2\pi a) + \overrightarrow{w_z^*}\ ; & \rho > 0.5 \end{cases}, \quad (19)$$

where ρ signifies the random number in [0, 1]. Thus, from the above equation, it is clear that the selection of the shrinking and the spiral movement is based on the random probability, as shown in the equation (19). On the other hand, during the exploration phase, the search agents update their positions randomly as

$$\overrightarrow{w_{z+1}} = \left|\overrightarrow{w_{rand}} - \overrightarrow{L}\cdot\overrightarrow{G}\right|. \quad (20)$$

During this mechanism, $\overrightarrow{L}$ exceeds one so that the global optimal search is executed in this phase. In other words, the global optimal solution is derived. The distance is computed as

$$\overrightarrow{G} = \overrightarrow{E}\cdot\overrightarrow{w_{rand}} - \overrightarrow{w_z}. \quad (21)$$

The modification of the WOA as experienced in equations (19) and (20) with the SGD is as follows. The update equation of SGD is given as

$$\overrightarrow{w_{z+1}} = \left(1 - \frac{1}{z}\right)\overrightarrow{w_z} + Y\cdot X, \quad (22)$$

where Y and X refer to the training sample indicating the total number of training samples and input feature vector. Rearranging the equation (22), we get

$$\overrightarrow{w_{z+1}} = \left(\frac{z-1}{z}\right)\overrightarrow{w_z} + Y\cdot X \quad (23)$$

$$\overrightarrow{w_z} = \left(\frac{z}{z-1}\right)\times\left[\overrightarrow{w_{z+1}} - Y\cdot X\right] \quad (24)$$

$$\overrightarrow{w_z} = \left(\frac{z}{z-1}\right)\times\overrightarrow{w_{z+1}} - Y\cdot X\times\left(\frac{z}{z-1}\right). \quad (25)$$

Modification of mathematical model—encircling phase: The standard equation of the WOA in encircling phase (equation (13)) is modified using the equation (25) as follows:

$$\overrightarrow{w_{z+1}} = \overrightarrow{w_z^*}\left[1 - \overrightarrow{L}\cdot\overrightarrow{E}\right] + \overrightarrow{L}\cdot\left(\frac{z}{z-1}\right) \times\overrightarrow{w_{z+1}} - \overrightarrow{L}\cdot\left(\frac{z}{z-1}\right)Y\cdot X \quad (26)$$

$$\overrightarrow{w_{z+1}} - \overrightarrow{L}\cdot\left(\frac{z}{z-1}\right)\times\overrightarrow{w_{z+1}} = \overrightarrow{w_z^*}\left[1 - \overrightarrow{L}\cdot\overrightarrow{E}\right] - \overrightarrow{L} \cdot\left(\frac{z}{z-1}\right)\times Y\cdot X \quad (27)$$

$$\left[1 - \overrightarrow{L}\cdot\left(\frac{z}{z-1}\right)\right]\times\overrightarrow{w_{z+1}} = \overrightarrow{w_z^*}\left[1 - \overrightarrow{L}\cdot\overrightarrow{E}\right] - \overrightarrow{L}\cdot\left(\frac{z}{z-1}\right)\times Y\cdot X \quad (28)$$

$$\left(\frac{z-1-\overrightarrow{L}\,z}{z-1}\right)\times\overrightarrow{w_{z+1}} = \overrightarrow{w_z^*}\left[1 - \overrightarrow{L}\cdot\overrightarrow{E}\right] - \overrightarrow{L}\cdot\left(\frac{z}{z-1}\right)\times Y\cdot X \quad (29)$$

$$\left(\frac{z\left(1-\overrightarrow{L}\right)-1}{z-1}\right)\times\overrightarrow{w_{z+1}} = \overrightarrow{w_z^*}\left[1 - \overrightarrow{L}\cdot\overrightarrow{E}\right] - \overrightarrow{L}\cdot\left(\frac{z}{z-1}\right)\times Y\cdot X \quad (30)$$

$$\overrightarrow{w_{z+1}} = \frac{z-1}{z(1-\overrightarrow{L})-1}\left\{\overrightarrow{w_z^*}\left[1 - \overrightarrow{L}\cdot\overrightarrow{E}\right] - \overrightarrow{L}\cdot\left(\frac{z}{z-1}\right)\times Y\cdot X\right\}. \quad (31)$$

It is noted from the above equation that the search agent's position is updated based on the coefficient vector, the iteration and the current best solution of the search agent.

Modification of mathematical model—exploitation phase: The standard equation of the WOA in exploitation phase [equation (18)] is modified using the equation (25).

$$\overrightarrow{w_{z+1}} = \overrightarrow{w_z^*}e^{da}\cdot\cos(2\pi a) - \left[\left(\frac{z}{z-1}\right) \times\overrightarrow{w_{z+1}} - Y\cdot X\times\left(\frac{z}{z-1}\right)\right].e^{da}\cdot\cos(2\pi a) + \overrightarrow{w_z^*} \quad (32)$$

$$\overrightarrow{w_{z+1}} + \left(\frac{z}{z-1}\right)\times\overrightarrow{w_{z+1}}\times e^{da}\cdot\cos(2\pi a) = \overrightarrow{w_z^*}e^{da} \cdot\cos(2\pi a) + Y\cdot X\times\left(\frac{z}{z-1}\right).e^{da}\cdot\cos(2\pi a) + \overrightarrow{w_z^*} \quad (33)$$

Downloaded from https://academic.oup.com/comjnl/advance-article-abstract/doi/10.1093/comjnl/bxz103/5627747 by Delhi Technological University user on 05 December 2019

$$\overrightarrow{w_{z+1}}\left\{1+\left(\frac{z}{z-1}\right)\times e^{da}\cdot\cos(2\pi a)\right\}=\overrightarrow{w_z^*}e^{da}\cdot\cos(2\pi a)+Y\cdot X\times\left(\frac{z}{z-1}\right).e^{da}\cdot\cos(2\pi a)+\overrightarrow{w_z^*} \quad (34)$$

$$\overrightarrow{w_{z+1}}=\frac{z-1}{(2z-1)\times e^{da}\cdot\cos(2\pi a)}\times\left\{\overrightarrow{w_z^*}e^{da}\cdot\cos(2\pi a)+Y\cdot X\times\left(\frac{z}{z-1}\right).e^{da}\cdot\cos(2\pi a)+\overrightarrow{w_z^*}\right\} \quad (35)$$

Modification of mathematical model—exploration phase: The update equation of the proposed algorithm during the random search is obtained through the integration of equation (25) in equations (20) and (21).

$$\overrightarrow{G}=\overrightarrow{E}\cdot\overrightarrow{w_{rand}}-\left(\frac{z}{z-1}\right)\times\overrightarrow{w_{z+1}}-Y\cdot X\times\left(\frac{z}{z-1}\right) \quad (36)$$

$$\overrightarrow{w_{z+1}}=\left|\overrightarrow{w_{rand}}-\overrightarrow{L}\cdot\left(\overrightarrow{E}\cdot\overrightarrow{w_{rand}}-\left(\frac{z}{z-1}\right)\times\overrightarrow{w_{z+1}}-Y\cdot X\times\left(\frac{z}{z-1}\right)\right)\right| \quad (37)$$

$$\overrightarrow{w_{z+1}}=\overrightarrow{w_{rand}}-\cdot\overrightarrow{L}\cdot\overrightarrow{E}\cdot\overrightarrow{w_{rand}}+\overrightarrow{L}\cdot\left(\frac{z}{z-1}\right)\times\overrightarrow{w_{z+1}}+\overrightarrow{L}.Y\cdot X\times\left(\frac{z}{z-1}\right) \quad (38)$$

$$\overrightarrow{w_{z+1}}\left[1-\overrightarrow{L}\cdot\left(\frac{z}{z-1}\right)\right]=\overrightarrow{w_{rand}}-\cdot\overrightarrow{L}\cdot\overrightarrow{E}\cdot\overrightarrow{w_{rand}}+\overrightarrow{L}.Y\cdot X\times\left(\frac{z}{z-1}\right) \quad (39)$$

$$\overrightarrow{w_{z+1}}=\frac{z-1}{2\left(1-\overrightarrow{L}\right)-1}\times\left\{\overrightarrow{w_{rand}}-\cdot\overrightarrow{L}\cdot\overrightarrow{E}\cdot\overrightarrow{w_{rand}}+\overrightarrow{L}.Y\cdot X\times\left(\frac{z}{z-1}\right)\right\} \quad (40)$$

It is understood that the search agent's position is updated based on the randomly chosen position of the search agent and in this step, the global optimal convergence is effective as the search agent widens its search outer the search area.

3.5.3. *Proposed SGD-based WOA*

The proposed SGD–WOA is deliberated in the following steps and algorithm 1 demonstrates the pseudocode of the proposed algorithm.

Population initialization: The total number of the H-whales in the search space is initialized as

$$w_z;\ \ (1\le z\le y), \quad (41)$$

where y refers to the total number of the search agents or H-whales in the search space.

Evaluate the fitness: The fitness of the solutions is evaluated for individual iteration to determine the best solution so that the position update of the search agents in the iteration follows the best solution. The fitness is evaluated based on the minimal error, and the solution corresponding to the minimal error is evaluated as the best solution. The error is determined as

$$MSE=\frac{1}{g}\left[\sum_{h=1}^{g}R_{h,t\arg et}-R_{h,est}\right]^2, \quad (42)$$

where g,$R_{h,est}$ and $R_{h,t\arg et}$ are the total frames of the video, estimated and target output of the classifier, respectively.

Update the optimization parameters: The optimization parameters, such as $\overrightarrow{L}$, $\overrightarrow{E}$, $\overrightarrow{e}$, ρ and a are updated at the end of each iteration so that there is an effective equilibrium between the exploration and the exploitation phases.

Update the position of the search agent: The search agents update their position based on the coefficient vector $\overrightarrow{L}$. Whenever the value of $\overrightarrow{L}$ lies below one, search agent's position is updated based on the equation in the encircling phase or through random search. On the other hand, whenever $\overrightarrow{L}$ exceeds 1, the search agent updates the position based on the spiral movement as in equation (40). To decide between the three phases, like encircling, exploration and exploitation, there is a probability. Whenever the probability ρ is less than 0.5 and $\overrightarrow{L}<1$, the search agent uses equation (31) to update the position. Whenever $\rho<0.5$ and $\overrightarrow{L}\ge 1$, the search agent encounters the random search procedure. Similarly, when $\rho\ge 0.5$, equation (35) is used to update the position of the search agent. The condition is

$$\overrightarrow{w_{z+1}}=\begin{cases}\frac{z-1}{z\left(1-\overrightarrow{L}\right)-1}\left\{\overrightarrow{w_z^*}\left[1-\overrightarrow{L}\cdot\overrightarrow{E}\right]-\overrightarrow{L}\cdot\left(\frac{z}{z-1}\right)\times Y\cdot X\right\}; & \rho<0.5\\ \frac{z-1}{(2z-1)\times e^{da}\cdot\cos(2\pi a)}\left\{\overrightarrow{w_z^*}e^{da}\cdot\cos(2\pi a)+Y\cdot X\times\left(\frac{z}{z-1}\right).e^{da}\cdot\cos(2\pi a)+\overrightarrow{w_z^*}\right\}; & \rho>0.5\end{cases}. \quad (43)$$

Downloaded from https://academic.oup.com/comjnl/advance-article-abstract/doi/10.1093/comjnl/bxz103/5627747 by Delhi Technological University user on 05 December 2019

Derive the optimal solution: The optimal search agent is decided based on the minimal fitness function.

Terminate: The update is continued for the maximal iterations to derive the global optimal solution.

Algorithm 1: SGD–WOA

Input: solution vector w_z
Output: optimal solution $\overrightarrow{w_z^*}$
  Initialization w_z;$(1 \leq z \leq y)$
  Evaluate the fitness
  While ($z <$ max *.no.ofiterations*)
    For all z
      Update $\overrightarrow{L}, \overrightarrow{E}, \overrightarrow{e}, \rho$ and$a$
        *If*1 ($\rho < 0.5$)
          *If*2($|L| < 1$)
            Update the position of the search agent by equation (31)
          Else *If*2 ($|L| \geq 1$)
            Choose the random search agent
            Update the position of search agent by equation (40)
          End *If*2
        Else *If*1 ($p \geq 0.5$)
          Update the position of search agent by equation (35)
        End *If*1
    End For
  Determine the best solution $\overrightarrow{w_z^*}$
  $z = z + 1$
  End While
  Return $\overrightarrow{w_z^*}$

Applications scenario: The CER finds applications in various fields such as in group testing of the user response to the video games, commercials where the products are tested in large scale automatically. It helps the autistics to interact with others and to spot the students struggling in classroom environment. It also helps to create virtual reality experiences. It adds security layer in malls, airports and in other public venues to detect malicious intent.

4. RESULTS AND DISCUSSION

The proposed SGD–WOA Deep CNN approach demonstrates the results in this section along with the comparative evaluation that proves the effectiveness of the proposed approach.

4.1. Experimental settings

The experimentation with MATLAB is performed using the real dataset that contains 70 videos collected from YouTube. These videos are collected by sending the user queries with seven emotions, such as angry, escape, fight, happy, normal, running/walking and violence. For each emotion, 10 videos are collected from YouTube and a total of 70 videos are collected for the recognition process under the category of normal and abnormal crowd videos. The proposed classifier is trained using the original features obtained from these videos and hence, upon the arrival of the test video, the object in the crowd video is located and the respective behaviors regarding the persons are recognized in the form of emotions or the seven classes described above.

4.2. Experimental analysis

This section discusses the emotions marked in the detection process of the proposed method using two videos, i.e. normal and abnormal videos. Figure 3 shows the experimental results obtained using the normal video with Fig. 3a–c depicting the input image, STI descriptor features of image and output of detection process for the image as 'Normal emotion'. Figure 4 shows the experimental results obtained using the abnormal video with Fig. 4a–f depicting the input image, STI descriptor features of image and output of detection process for the image as running cum walking emotion, angry emotion, fight emotion, happy emotion, escape emotion and violence emotion, respectively.

4.3. Performance metrics

The evaluation metrics [32, 33, 34], such as sensitivity (SENS), specificity (SPEC) and accuracy (ACCU), analyze the performance of the proposed SGD–WOA Deep CNN involved in the CER. The mathematical expression of the metrics is given below.

4.3.1. *Sensitivity*

It measures the true positives correctly detected using the proposed approach and it is defined as

$$Sensitivity = \frac{TP}{TP + FP}, \tag{44}$$

where TP is the true positive and TN denotes the true negative.

4.3.2. *Specificity*

It denotes the true negatives correctly identified using the proposed approach and it is defined as

$$Specficity = \frac{TN}{TN + FP}, \tag{45}$$

where TN is the true negative and FP denotes the false positive.

Downloaded from https://academic.oup.com/comjnl/advance-article-abstract/doi/10.1093/comjnl/bxz103/5627747 by Delhi Technological University user on 05 December 2019



FIGURE 3. Experiment using normal video; (a) input image, (b) output image by STI descriptor and (c) detected output image using the classifier as 'Normal emotion'.

4.3.3. *Accuracy*

It denotes the measure of the closeness of the proposed approach in detecting the crowd behavior and is given as

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \quad (46)$$

4.4. Performance analysis

The section discusses the performance analysis of the proposed approach by varying the sizes of the conv layers ($32 \times 32 \times 3$) and filter size ($3 \times 3 \times 3$) using the performance metrics.

4.4.1. *Performance analysis based on width of conv layers*

The performance of the proposed SGD–WOA Deep CNN approach is analyzed based on the metrics, such as SENS, SPEC and ACCU and Fig. 5 exhibits the analysis based on the varying width of *conv* layer. Figure 5a exhibits the analysis between SENS and training percentage. The width of the conv layers considered for the analysis includes 4, 8, 12, 16, 20 and 24, respectively. When the width of the *conv* layers is 4, the SENS of the proposed approach is 0.9188, and for the width of *conv* layers 8, 12, 16, 20 and 24, the SENS of the proposed approach is 0.9273, 0.9273, 0.9273, 0.9273 and 0.9316, respectively, for the training percentage of 0.4. The SENS of the proposed approach is 0.97 for all the widths of *conv* layers when the training percentage is 0.8.

Figure 5b exhibits the analysis between SPEC and training percentage. The width considered for the analysis includes 4, 8, 12, 16, 20 and 24, respectively. When the width of the *conv* layers is 4, the SPEC of the proposed approach is 0.9898, and for the width of *conv* layers 8, 12, 16, 20 and 24, the SPEC of the proposed approach is 0.9908, 0.9908, 0.9908, 0.9908 and 0.9914, respectively, for the training percentage of 0.4. The SPEC of the proposed approach is 0.97 for all the widths of *conv* layers when the training percentage is 0.8.

Figure 5C exhibits the analysis between ACCU and training percentage. The width considered for the analysis includes 4, 8, 12, 16, 20 and 24, respectively. When the width of the *conv* layers is 4, the ACCU of the proposed approach is 0.9477, and for the width of *conv* layers 8, 12, 16, 20 and 24, the ACCU of the proposed approach is 0.9501, 0.9501, 0.9501, 0.9501 and 0.9512, respectively, for the training percentage of 0.4. The ACCU of the proposed approach is 0.97 for all the widths of *Conv* layers when the training percentage is of 0.8.

4.4.2. *Performance analysis based on the filter sizes*

The performance of the proposed SGD–WOA Deep CNN approach is analyzed based on the metrics, such as SENS, SPEC and ACCU and Fig. 6 exhibits the analysis based on the varying filter size. Figure 6a exhibits the analysis between SENS and training percentage. The filter sizes considered for the analysis include 25, 30, 75, 100, 125 and 150, respectively. When the filter size is 25, the SENS of the proposed approach is 0.9102, and for the filter sizes, 30, 75, 100, 125 and 150, the SENS of the proposed approach is 0.9145, 0.9188, 0.9230, 0.9273 and 0.9401, respectively, for the training percentage of 0.4. The SENS of the proposed approach is 0.97 for all the filter sizes when the training percentage is 0.8.

Figure 6b exhibits the analysis between SPEC and training percentage. When the filter size is 25, the SPEC of the proposed approach is 0.9887, and for the filter sizes, 30, 75, 100, 125 and 150, the SPEC of the proposed approach is 0.9892, 0.9898, 0.9903, 0.9908 and 0.9924, respectively, for the training percentage of 0.4. The SPEC of the proposed approach is 0.97 for all the filter sizes when the training percentage is 0.8.

Figure 6C exhibits the analysis between ACCU and training percentage. The width considered for the analysis includes 25, 30, 75, 100, 125 and 150, respectively. When the filter size is 25, the ACCU of the proposed approach is 0.9454, and for the filter sizes, 30, 75, 100, 125 and 150, the ACCU of the proposed approach is 0.9465, 0.9477, 0.9489, 0.9501 and 0.9536, respectively, for the training percentage of 0.4. The ACCU of the proposed method is 0.97 for the filter size 25 when the training percentage is 0.8.

4.5. Competing approaches

The approaches used for the comparison include DBN [31], deep convolutional neural networks-long short-term memory (DCNN-LSTM) [30], linear SVM [17], radial basis function Kernel-based SVM (RBF Kernel-based SVM) [14], GMM [18] and Deep CNN [25] for the effective comparison with the proposed SGD–WOA Deep CNN approach.

Downloaded from https://academic.oup.com/comjnl/advance-article-abstract/doi/10.1093/comjnl/bxz103/5627747 by Delhi Technological University user on 05 December 2019

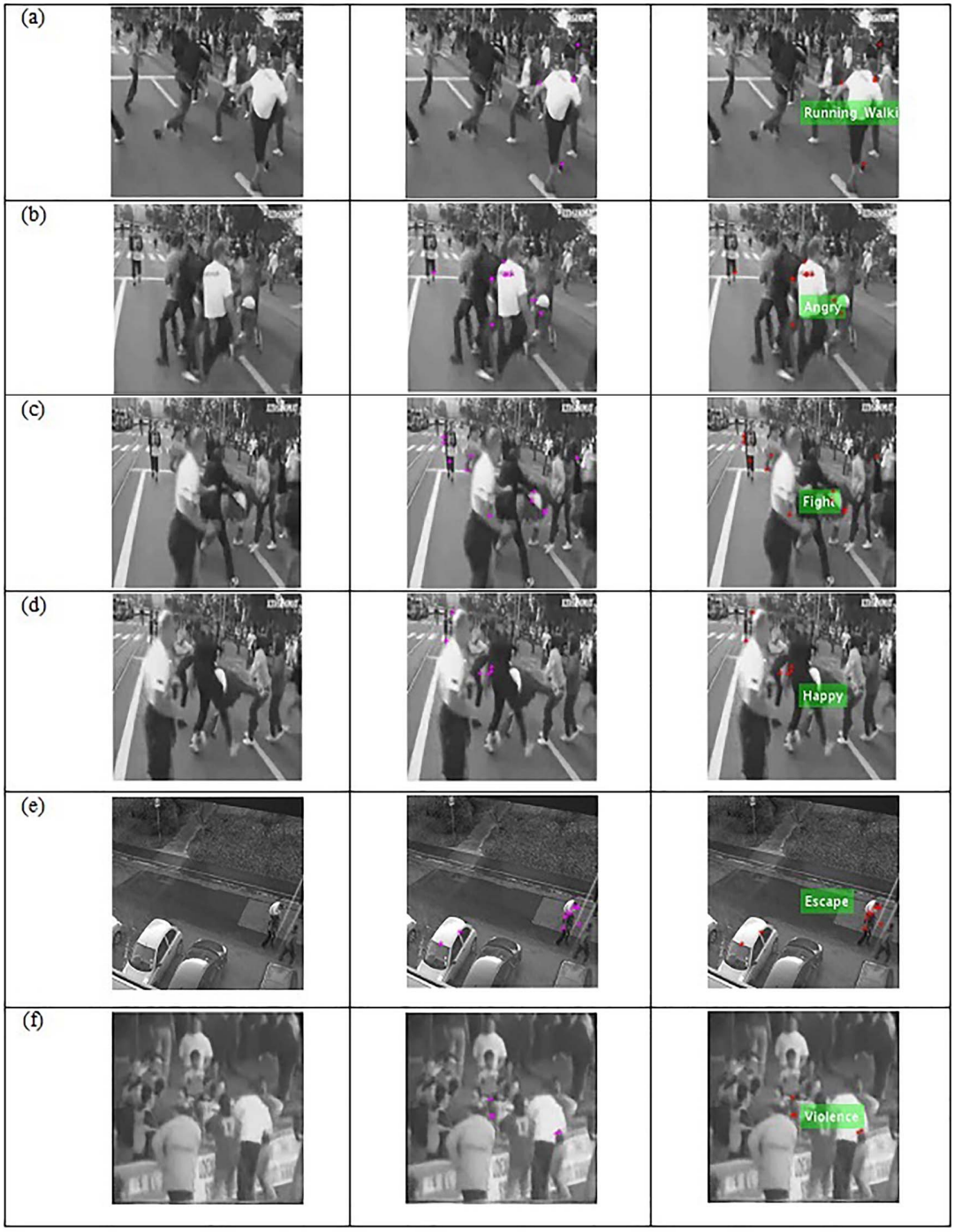


FIGURE 4. Experiment using abnormal videos with first image as an input image, second image as the output from STI descriptor and third image as the detected output image using the classifier; (a) running cum walking emotion, (b) angry emotion, (c) fight emotion, (d) happy emotion, (e) escape emotion and (f) violence emotion.

Downloaded from https://academic.oup.com/comjnl/advance-article-abstract/doi/10.1093/comjnl/bxz103/5627747 by Delhi Technological University user on 05 December 2019

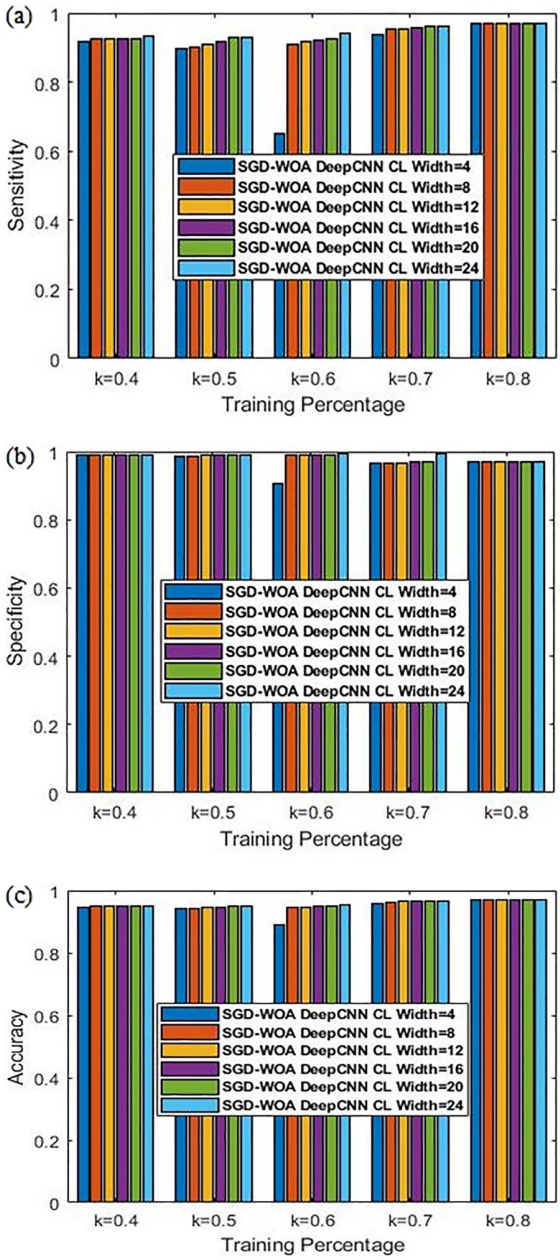


FIGURE 5. Performance analysis by varying the width of the conv layers; (a) SENS, (b) SPEC and (c) ACCU.

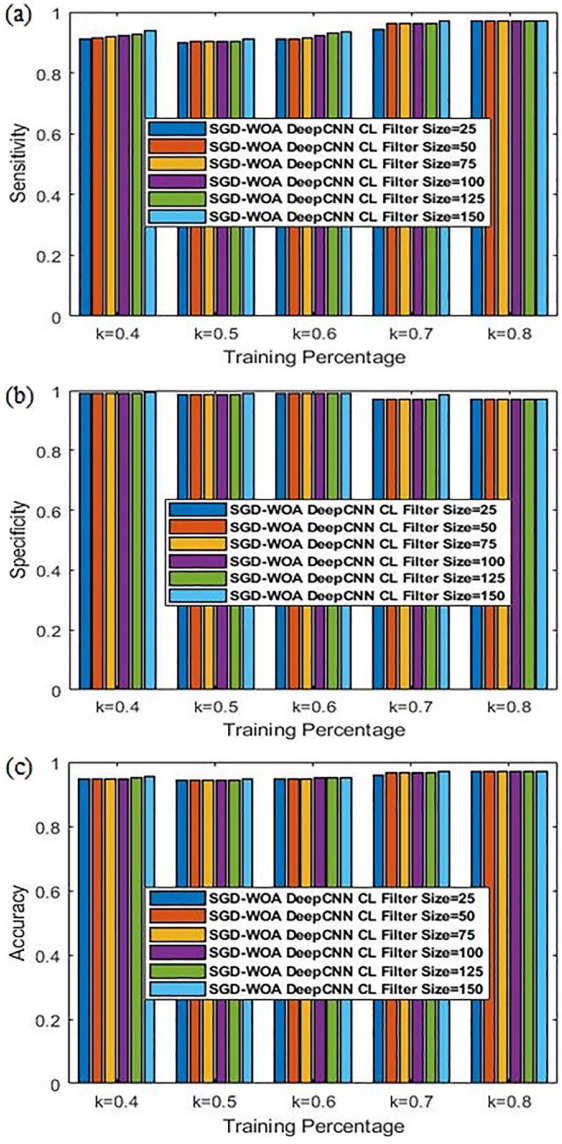


FIGURE 6. Performance analysis by varying the filter sizes; (a) SENS, (b) SPEC and (c) ACCU.

4.6. Comparative evaluation

The section demonstrates the comparative evaluation of the approaches based on the performance metrics depending on the training percentage and the K-fold values.

4.6.1. Based on training percentage

The comparative evaluation of the approaches is performed based on the performance metrics, such as SENS, SPEC and ACCU and Fig. 7 exhibits the analysis based on varying training percentages. Figure 7a exhibits the analysis between SENS and training percentage. The SENS of the approaches DBN, DCNN-LSTM, Linear SVM, RBF Kernel-based SVM, GMM, Deep CNN and SGD–WOA Deep CNN is 0.6, 0.8633, 0.6642, 0.9059, 0.9059, 0.9486 and 0.9529, respectively, for the training percentage of 0.4. Thus, the proposed approach acquired a maximal value of SENS as compared to the existing approaches.

Figure 7b exhibits the analysis between SPEC and training percentage. The SPEC of the approaches DBN, DCNN-LSTM, Linear SVM, RBF Kernel-based SVM, GMM, Deep

Downloaded from https://academic.oup.com/comjnl/advance-article-abstract/doi/10.1093/comjnl/bxz103/5627747 by Delhi Technological University user on 05 December 2019

CNN, and SGD–WOA Deep CNN is 0.8419, 0.8633, 0.9561, 0.9807, 0.9882, 0.9882 and 0.9941, respectively, for the training percentage of 0.4. Thus, the proposed approach acquired a maximal value of SPEC as compared to the existing approaches.

Figure 7C exhibits the analysis between ACCU and training percentage. The ACCU of the approaches DBN, DCNN-LSTM, Linear SVM, RBF Kernel-based SVM, GMM, Deep CNN and SGD–WOA Deep CNN is 0.6269, 0.8633, 0.9010, 0.9442, 0.9442, 0.9565 and 0.9571, respectively, for the training percentage of 0.4. Thus, the proposed approach acquired a maximal value of ACCU as compared to the existing approaches.

4.6.2. *Based on K-fold*

The comparative analysis of the approaches is analyzed based on SENS, SPEC and ACCU and Fig. 8 exhibits the analysis based on varying K-fold values. Figure 8a exhibits the analysis between SENS and K-fold values. The SENS of the approaches DBN, DCNN-LSTM, Linear SVM, RBF Kernel-based SVM, GMM, Deep CNN and SGD–WOA Deep CNN is 0.6, 0.9174, 0.7901, 0.9425, 0.9496, 0.9665 and 0.9675, respectively, for the K-fold of 5. Thus, the proposed approach acquired a maximal value of SENS as compared to the existing approaches.

Figure 8b exhibits the analysis between SPEC and K-fold values. The SPEC of DBN, DCNN-LSTM, Linear SVM, RBF Kernel-based SVM, GMM, Deep CNN and SGD–WOA Deep CNN is 0.9411, 0.9685, 0.9422, 0.9688, 0.9737, 0.9851 and 0.9936, respectively, for the K-fold of 5. The proposed approach acquired a maximal value of SPEC as compared to the existing approaches.

Figure 8C exhibits the analysis between ACCU and K-fold values. The ACCU of the approaches DBN, DCNN-LSTM, Linear SVM, RBF Kernel-based SVM, GMM, Deep CNN and SGD–WOA Deep CNN is 0.6715, 0.9350, 0.8596, 0.9545, 0.9587, 0.9685 and 0.9693, respectively, for the K-fold of 5. The proposed approach acquired a maximal value of ACCU as compared to the existing approaches.

Figure 9 exhibits the ROC analysis of the approaches and it is observed that for the minimum FPR of 0.1, the TPR of the proposed SGD–WOA Deep CNN approach is maximal of 0.9523, whereas for the approaches, DBN, DCNN-LSTM, Linear SVM, RBF Kernel-based SVM, GMM and Deep CNN, the TPR is 0.8399, 0.6, 0.7877, 0.9370, 0.9487 and 0.9523, respectively.

4.7. Comparative discussion

The comparative evaluation of the approaches based on the K-fold is discussed in Table 1. The approaches, like DBN, DCNN-LSTM, Linear SVM, RBF Kernel-based SVM, GMM, Deep CNN and SGD–WOA Deep CNN, acquired the SPEC of 0.9104, 0.9674, 0.9605, 0.9682, 0.9698, 0.9754 and 0.9807, respectively, when the K-fold is 10. Likewise, the ACCU

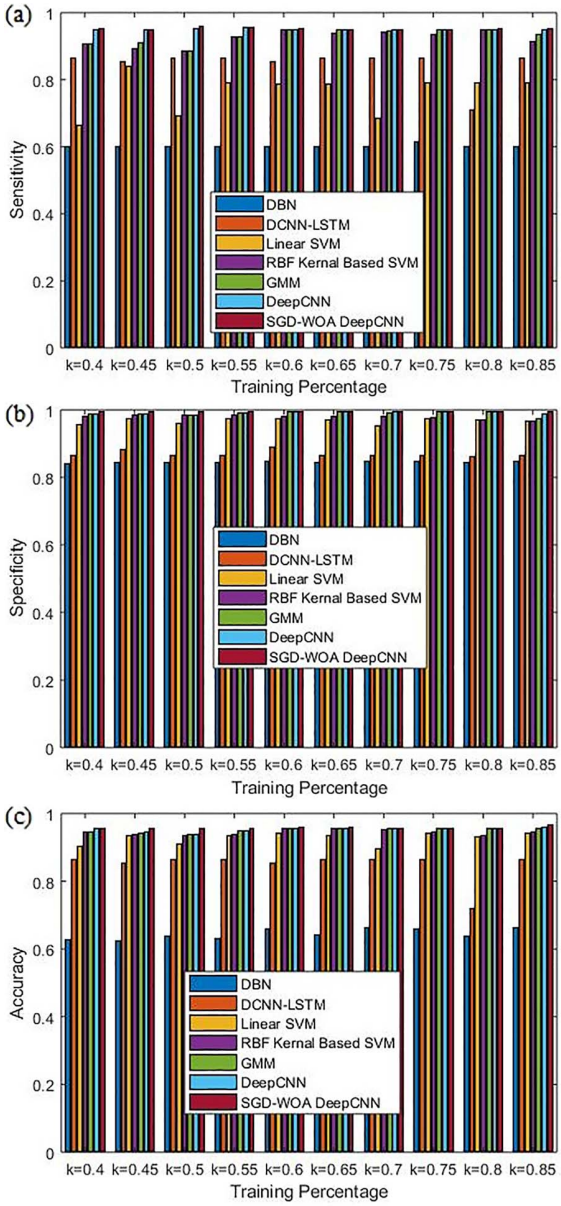


FIGURE 7. Comparative analysis by varying the training percentage; (a) SENS, (b) SPEC and (c) ACCU.

of approaches like DBN, DCNN-LSTM, Linear SVM, RBF Kernel-based SVM, GMM, Deep CNN and SGD–WOA Deep CNN is 0.6, 0.8582, 0.9204, 0.9602, 0.9630, 0.9679 and 0.9693, respectively, for the K-fold 10. The SENS of the proposed approach at the K-fold 10 is 0.9675, whereas SENS for the approaches DBN, DCNN-LSTM, Linear SVM, RBF Kernel-based SVM, GMM and Deep CNN is 0.6, 0.8183, 0.8137, 0.9448, 0.9514, 0.9659 and 0.9675, respectively.

The evaluation based on the training percentage is demonstrated in Table 2. The ACCU of DBN, DCNN-LSTM, Lin-

Downloaded from https://academic.oup.com/comjnl/advance-article-abstract/doi/10.1093/comjnl/bxz103/5627747 by Delhi Technological University user on 05 December 2019

TABLE 1. Comparative evaluation based on K-fold.

Methods	Based on the K-fold			
	Sensitivity	Specificity	Accuracy	Computational time (secs)
DBN [31]	0.6	0.9104	0.6	12
DCNN-LSTM [30]	0.8183	0.9674	0.8582	10
Linear SVM [17]	0.8137	0.9605	0.9204	11.7
RBF Kernel-based SVM [14]	0.9448	0.9682	0.9602	8
GMM [18]	0.9514	0.9698	0.9630	11
Deep CNN [25]	0.9659	0.9754	0.9679	9
SGD–WOA Deep CNN [proposed]	0.9675	0.9807	0.9693	5

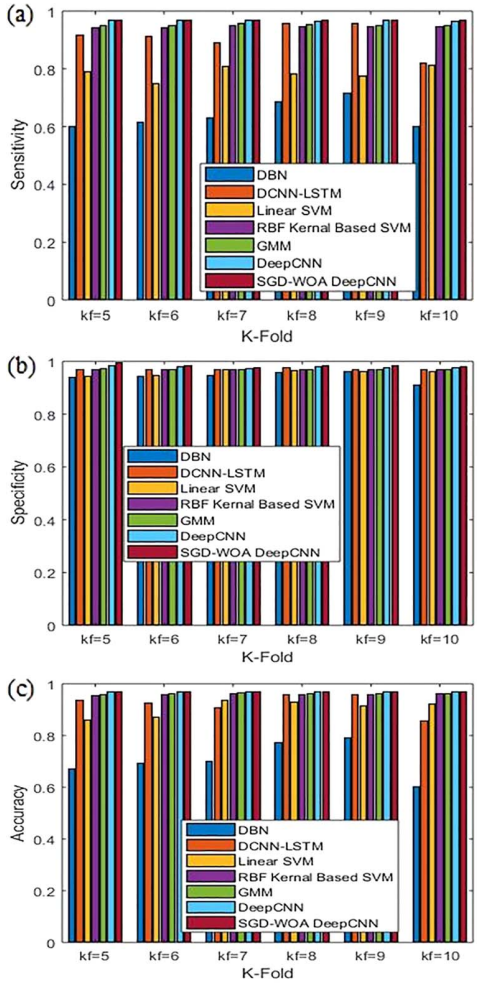


FIGURE 8. Performance analysis by varying the K-fold; (a) SENS, (b) SPEC and (c) ACCU.

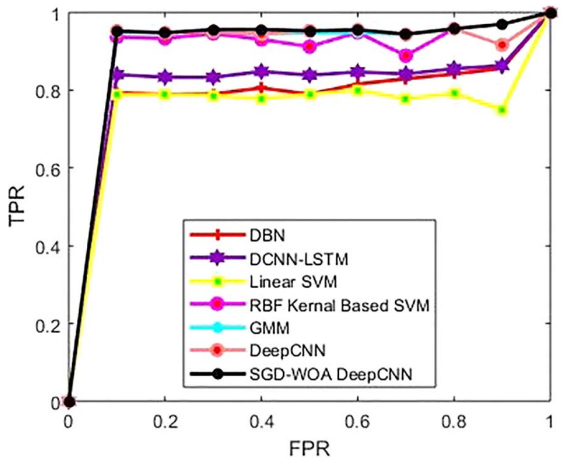


FIGURE 9. ROC analysis.

ear SVM, RBF Kernel-based SVM, GMM, Deep CNN and SGD–WOA Deep CNN is 0.6624, 0.8633, 0.9420, 0.9466, 0.9559, 0.9606 and 0.9653, respectively. The SENS values of DBN, DCNN-LSTM, Linear SVM, RBF Kernel-based SVM, GMM, Deep CNN and SGD–WOA Deep CNN are 0.6, 0.8633, 0.7895, 0.9123, 0.9360, 0.9474 and 0.9530, respectively. The SPEC of DBN, DCNN-LSTM, Linear SVM, RBF Kernel-based SVM, GMM, Deep CNN and SGD–WOA Deep CNN is 0.8474, 0.8633, 0.9659, 0.9679, 0.9744, 0.9893 and 0.9936, respectively. The computation time of DBN, DCNN-LSTM, Linear SVM, RBF Kernel-based SVM, GMM, Deep CNN and SGD–WOA Deep CNN are 12 s, 10s, 11.7 s, 8 s, 11 s, 9 s and 5 s, respectively. The computational time of the proposed approach is lesser than the other approaches, thus improving the efficiency of the system.

5. CONCLUSION

The automatic CER is progressed in this paper using the SGD–WOA Deep CNN classifier. The traditional approaches employed the voice, face and so on, to recognize the behaviors of the crowd, which are ineffective as they are incapable of dealing with the dynamics of the objects in the video. Initially,

Downloaded from https://academic.oup.com/comjnl/advance-article-abstract/doi/10.1093/comjnl/bxz103/5627747 by Delhi Technological University user on 05 December 2019

TABLE 2. Comparative evaluation based on training percentage.

	Based on the training percentage		
Methods	Sensitivity	Specificity	Accuracy
DBN [31]	0.6	0.8474	0.6624
DCNN-LSTM [30]	0.8633	0.8633	0.8633
Linear SVM [17]	0.7895	0.9659	0.9420
RBF Kernel-based SVM [14]	0.9123	0.9679	0.9466
GMM [18]	0.9360	0.9744	0.9559
Deep CNN [25]	0.9474	0.9893	0.9606
SGD–WOA Deep CNN [proposed]	0.9530	0.9936	0.9653

in the proposed approach of automatic recognition, the key frames from the input video are chosen in such a way that the frames with significant features are chosen. For the selection of the key frames, the frames in the video are subjected to the WT and, using the Bhattacharya distance, the key frames are chosen. After the selection of the key frames, the features from the frames are chosen using the STI descriptor, and the features are fed as input to the deep CNN classifier. The classifier is tuned optimally using the proposed SGD–WOA, which is the unification of SGD and WOA. The analysis of the approach using the datasets proves that the proposed approach acquired a better ACCU of 0.9693, SENS of 0.9675 and SPEC of 0.9936.

REFERENCES

[1] Bhargava, N. and Chaudhuri, S. (2017) An integrated approach to crowd video analysis: from tracking to multi-level activity recognition. 1–10. https://arxiv.org, CoRR abs/1710.11087.

[2] Zhan, B., Monekosso, D.N., Remagnino, P., Velastin, S.A. and Xu, L.-Q. (2008) Crowd analysis: a survey. *Mach. Vis. Appl.*, 19, 345–357.

[3] Saxena, S., Bremond, F., Thonnat, M. and Ma, R. (2004) Crowd behavior recognition for video surveillance. *Proc. Int. Conf. Advanced Concepts for Intelligent Vision Systems*, pp. 970–981. Juan-Les-Pins, France.

[4] Baig, M.W., Barakova, E.I., Marcenaro, L., Regazzoni, C.S. and Rauterberg, M. (2014) Bio-inspired probabilistic model for crowd emotion detection. In *Proc. IEEE Int. Joint Conf. Neural Networks*, pp. 3966–3973. Beijing, China.

[5] Bower, G.H. and Cohen, P.R. (1982) Emotional influences in memory and thinking: data and theory. In *Affect and Cognition: 17th Annual Carnegie Mellon Symposium on Cognition*, pp. 291–331.

[6] Baig, M.W. *et al.* (2015) Perception of emotions from crowd dynamics. In *Proc. IEEE Int. Conf. Digital Signal Processing*, pp. 703–707. Singapore.

[7] Fabrega, H. (2000) The feeling of what happens: body and emotion in the making of consciousness. *Psychiatr. Serv.*, doi:10.1176/appp.ps.41.12.1579.

[8] Villon, O. and Lisetti, C. (2006) A user-modeling approach to build user's psycho-physiological maps of emotions using bio-sensors. *Proc. 15th IEEE Int. Symposium on Robot and Human Interactive Communications*, pp. 269–276. Hatfield, UK.

[9] Liu, M., Wang, R., Li, S., Shan, S., Huang, Z. and Chen, X. (2014) Combining multiple kernel methods on Riemannian manifold for emotion recognition in the wild. In *Proc. 16th Int. Conf. Multimodal Interaction*, pp. 494–501. Istanbul, Turkey.

[10] Prasomphan, S. (2015) Improvement of speech emotion recognition with neural network classifier by using speech spectrogram. In *Proc. 2015 Int. Conf. on Systems, Signals and Image Processing*, pp. 73–76. London, UK.

[11] Ingale, B.A. and Chaudhari, D.S. (2012) Speech emotion recognition using hidden Markov model and support vector machine. *Int. J. Adv. Eng. Res. Stud.*, 1, 316–318.

[12] Gaikwad, K. and Narawade, V. (2012) HMM classifier for human activity recognition. *Comp. Sci. Eng.*, 2, 27–36.

[13] Patwardhan, A.S. (2016) Walking, lifting, standing activity recognition using probabilistic networks. *Int. Res. J. Eng. Technol.*, 3, 667–670.

[14] Patwardhan, A. (2016) Edge based grid super-imposition for crowd emotion recognition. *Int. Res. J. Eng. Technol.*, 3, 459–463.

[15] Milton, A., Roy, S.S. and Selvi, S.T. (2013) SVM scheme for speech emotion recognition using MFCC feature. *Int. J. Comp. Appl.*, 69, 34–39.

[16] Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K. and Mahjoub, M.A. (2018) Speech emotion recognition: methods and cases study. In *Proc. 10th Int. Conf. Agents and Artificial Intelligence*, Vol. 2, pp. 175–182. Funchal, Madeira, Portugal.

[17] Rabiee, H., Haddadnia, J., Mousavi, H., Nabi, M., Murino, V. and Sebe, N. (2016) Crowd behavior representation: an attribute-based approach. *Springerplus*, 5, 2214 and 1179.

[18] Baig, M.W., Barakova, E.I., Marcenaro, L., Rauterberg, M. and Regazzoni, C.S. (2014) Crowd emotion detection using dynamic probabilistic models. In *From Animal to Animats 13: Proc. 13th Int. Conf. Simulation of Adaptive Behavior*, pp. 328–337. Castellon, Spain.

[19] He, X., Barakova, E.I., Regazzoni, C.S., Marcenaro, L., Urizar, O.J., Rauterberg, M. and Baig, M.S. (2016) A hierarchical Bayesian model for crowd emotions. *Front. Comput. Neurosci.*, 10.

[20] He, X. and Zhang, W. (2018) Emotion recognition by assisted learning with convolutional neural networks. *Neurocomputing*, 291, 187–194.

[21] Duncan, D., Shine, G., and English, C. (2016) *Facial emotion recognition in real time*. cs231n.stanford.edu./reports/2016/pdfs/022-report.pdf.

[22] Wagner, J., Kim, J. and André, E. (2005) From physiological signals to emotions: implementing and comparing selected methods for feature extraction and classification. In *Proc. 2005 IEEE Int. Conf. Multimedia and Expo*, pp. 940–943. Amsterdam, Netherlands.

Downloaded from https://academic.oup.com/comjnl/advance-article-abstract/doi/10.1093/comjnl/bxz103/5627747 by Delhi Technological University user on 05 December 2019

[23] Schuldt, C., Laptev, I. and Caputo, B. (2004) Recognizing human actions: a local SVM approach. In *Proc. 17th Int. Conf. Pattern Recognition*, Vol. 3, pp. 32–36. Cambridge, UK.

[24] Babu, G.S., Zhao, P. and Li, X.-L. (2016) Deep convolutional neural network based regression approach for estimation of remaining useful life. In *Proc. Int. Conf. Database Systems for Advanced Applications*, pp. 214–228. Dallas, TX, USA.

[25] Tu, F., Yin, S., Ouyang, P., Tang, S., Liu, L. and Wei, S. (2017) Deep convolutional neural network architecture with reconfigurable computation patterns. *IEEE Trans. Very Large Scale Integr. VLSI Syst.*, 25, 2220–2233.

[26] Yang, J. and Yang, G. (2018) Modified convolutional neural network based on dropout and the stochastic gradient descent optimizer. *Algorithms*, 11, 1–15.

[27] Mirjalili, S. and Lewis, A. (2016) The whale optimization algorithm. *Adv. Eng. Softw.*, 95, 51–67.

[28] Liu, Z., Xie, Q., Wu, M., Cao, W.-H., Mei, Y. and Mao, J.-W. (2018) Speech emotion recognition based on an improved brain emotion learning model. *Neurocomputing*, 309, 145–156.

[29] Yan, J., Zheng, W., Cui, Z., Tang, C., Zhang, T. and Zong, Y. (2018) Multi-cue fusion for emotion recognition in the wild. *Neurocomputing*, 309, 27–35.

[30] Wei, Q., Zhao, Y., Xu, Q., Li, L., He, J., Yu, L. and Sun, B. (2017) A new deep-learning framework for group emotion recognition. In *Proc. 19th ACM Int. Conf. Multimodal Interaction*, pp. 587–592. Glasgow, UK.

[31] Uddin, M.Z., Hassan, M.M., Almogren, A., Alamri, A., Alrubaian, M. and Fortino, G. (2017) Facial expression recognition utilizing local direction-based robust features and deep belief network. *IEEE Access*, 5, 4525–4536.

[32] Ratre, A. and Pankajakshan, V. (2018) Tucker tensor decomposition based tracking and Gaussian mixture model for anomaly localisation and detection in surveillance videos. *IET Comp. Vis.*, 12, 933–940.

[33] Ratre, A. and Pankajakshan, V. (2018) Tucker visual search-based hybrid tracking model and fractional Kohonen self-organizing map for anomaly localization and detection in surveillance videos. *Imaging Sci. J.*, 66, 195–210.

[34] Ratre, A. (2018) Taylor series-based compressive approach and firefly support vector neural network for tracking and anomaly detection in crowded videos. *J. Eng. Res.*, in press.

[35] Ratre, A. (2019) Whale optimization-based particle filter for multi-object tracking in surveillance videos. In *1st IEEE Int. Conf. Signal Processing, VLSI and Communication Engineering*, pp. 28–30. Delhi, India, in press.

Downloaded from https://academic.oup.com/comjnl/advance-article-abstract/doi/10.1093/comjnl/bxz103/5627747 by Delhi Technological University user on 05 December 2019

Structural, Physical, Electrical and Dielectric Properties of Magnetic Glasses: $x\mathrm{Fe_2O_3} + (30 - x)\mathrm{V_2O_5} + 30\mathrm{Na_2O} + 40\mathrm{B_2O_3}$ with $x = 0$ to 15

P. Chand[a,b,*], L. Kumar[b], A. Yadav[c] and S. Khasa[d]

[a]Department of Physics, Netaji Subhas University of Technology, Sector 3, Dwarka, New Delhi — 110078, India

[b]Material Science, Department of Physics, SRM University, Delhi-NCR Sonepat — 131029, Haryana, India

[c]Department of Applied Physics, Delhi Technological University, Delhi — 110042, India

[d]Materials Research Laboratory, Department of Physics, DCR University of Science and Technology, Sonepat — 131039, Haryana, India

(Received April 14, 2019; revised version October 1, 2019, in final form November 4, 2019)

Magnetic glasses compositions: $x\mathrm{Fe_2O_3} + (30 - x)\mathrm{V_2O_5} + 30\mathrm{Na_2O} + 40\mathrm{B_2O_3}$ (x = 0, 2, 3, 5, 7, 10, and 15), abbreviated as FVNBx, are prepared with conventional melt-quenching technique, and the structural, physical, and electrical properties of compositions are investigated. The amorphous nature of the FVNBx glasses is confirmed using X-ray diffraction technique. It is found that the density of the glasses increases with x. The molar volumes calculations are based on the molar mass and the density data, which are then correlated with the structural changes. The decrease in the theoretical optical basicity shows marginal increase in the covalence of oxygen, when the $\mathrm{V_2O_5}$ content is replaced with $\mathrm{Fe_2O_3}$. The Fourier transform infrared spectroscopy that is carried out in the mid-IR region, shows the absence of the six-atomic ($\mathrm{B_3O_3}$) boroxol rings. In this paper, the IR absorption bands corresponding to the combined contributions of tri and tetra borate vibrations are identified. The distinct IR spectra of the samples with different x indicates that the iron acts as a strong glass modifier, and brings only noticeable microstructural changes in the glass structures. $\mathrm{Fe_2O_3}$ is believed to act as both glass former and glass modifier. The increase of DC electrical conductivity with temperature indicates a semiconducting nature of the glasses. The replacement of $\mathrm{V_2O_5}$ by $\mathrm{Fe_2O_3}$, even by 1 mol%, decreases the DC conductivity by an order of magnitude, and increases the activation energy by a factor of two. Composition, frequency, and temperature dependence of different dielectric parameters, namely dielectric constant ε', dielectric loss ε'', loss tangent $\tan(\delta)$, electrical modulus M^*, etc., are also analyzed.

DOI: 10.12693/APhysPolA.136.897

PACS/topics: soda-iron-boro-vanadate glasses, FTIR, dielectric, electrical conductivity, electrical modulus, loss tangent

1. Introduction

Glasses form a very distinct class of materials, which become important due to its countless applications. Traditional use of glasses includes containers, windows, lamps and optical components etc. More specific use includes applications in communication, electronics, and containers for radioactive waste. Any development that demonstrates the importance, versatility, and promise of glasses for newer and newer use, makes these materials very attractive for researchers. The glasses are prepared by almost limitless different chemical compositions, which give rise to different properties. Borate glasses are known for lower temperature fabrication, variety of chemical compositions, and for their structural peculiarities. This is due to their dielectric, optical, magnetic, and super ionic properties [1–5].

The structural information is usually obtained by X-ray and infrared investigations. The thermal study provides information about the stability and the glass transition temperature. The electrical properties are investigated by the dielectric measurements and the AC and DC electric conductivities.

From previous studies on glasses, it is well understood that glasses can exhibit both electronic conduction as well as polaronic one. This latter conduction is directly dependent on the local structure and carrier concentration in the material under study. Jozwiak et al. had investigated the electrical and thermal properties of lithium–iron–phosphate (LFP) glasses and found that glasses with olivine-like local order exhibit mixed polaronic–ionic conductivity with predominance of the polaronic component [6]. They have also shown the possibility of enhancement in total electrical conductivity by the creation of nanocrystallites in the lithium–iron–phosphate glass. Garbarczyk et al. [7] have studied electrical properties of mixed conductive silver vanadate–phosphate glasses and observed the ion-polaron effect (IPE) due to electrostatic attraction

*corresponding author; e-mail: prem@iitk.ac.in

between electrons/polarons and Ag^+ ions. This IPE effect results in decrease of total electrical conductivity at specifically given concentrations of both components. Kamitsos et al. have studied the lithium sulfate based fast ionic conducting borate glasses by vibrational and electric transport measurements. They found an evidence for two broad distributions of cation-hosting environments [8, 9]. Bih et al. studied the electronic and ionic conductivity of glasses of the Li_2O–MoO_3–P_2O_5 system [10]. Recently, Langar et al. investigated the conductivity and dielectric behaviour of $NaPO_3$–ZnO–V_2O_5 glasses [11]. Mekki et al. have investigated the structure and magnetic properties of vanadium–sodium–silicate glasses [12]. Bhagat studied the dielectric relaxation in Na-doped BiPbSrCaCuO glasses [13]. Dalba et al. have investigated the short range order in borate glasses by X-ray absorption spectroscopy [14]. An electrothermal model for high field conduction and switching phenomena in Te_2O_2–V_2O_5 glasses was suggested by Montani et al. [15].

Our recent studies concern various properties of different vanadium doped borate glasses [16–18]. Kashif et al. have studied the effect of alkali content on AC conductivity of borate glasses containing two transition metals [19]. Borate glasses are generally insulating in nature and the addition of transition metal oxides, such as Fe_2O_3 and V_2O_5, makes them dilute magnetic in nature and electrically semiconducting. These mentioned properties create possibilities for preparing such glasses having potential applications in optical and electronic memory switching, cathode materials for making solid state devices, and optical fibres. Addition of alkali metal oxides to borate network results in the formation of several borate groups, which lead to tightly organized structures in an intermediate range order. Alkali borate glasses having transition metal ions could open up new possibilities to extend the properties of these materials.

Mekki et al. reported that in vanadium–sodium–silicate glasses only 2% of total vanadium was found in magnetic V^{4+} oxidation state, most of it was found to exist in the non-magnetic V^{5+} oxidation state [12]. Saetova et al. have studied the Li_2O–V_2O_5–B_2O_3 glasses and, similarly to borate case, found that vanadium exists predominantly in non-magnetic V^{5+} valence whereas the magnetic V^{4+} valence is in minority, while the electrical conduction is mainly electronic [20].

There exists such a conviction that the increase in the average spacing between the polaron hoping sites inhibits the mobility of small polarons. Thus, observed decrease in the electrical conductivity and increase in the activation energy might be affected by increase in energy hopping. Network modification reveals the formation of non-bridging oxygens (NBOs), and converting basic structural units of borate and vanadate into charged units. The concentration of the NBOs determines ease of bond reconstruction and hence the viscosity. The microstructural information is conventionally obtained through characteristic Fourier transform infrared (FTIR) spectra, X-ray diffraction (XRD), and other spectroscopic techniques. Presence of iron in the glasses ensures that these glasses are magnetic since both Fe^{2+} and Fe^{3+} oxidation states are magnetic, whereas vanadium can exist entirely in non-magnetic V^{5+} oxidation state.

In the present paper, we have attempted to study the magnetic soda–boro–vanadate–iron glasses of the composition: $x\mathrm{Fe_2O_3} + (30 - x)\mathrm{V_2O_5}{+}30\mathrm{Na_2O} + 40\mathrm{B_2O_3}$ (FVNBx, where $x = 0, 2, \ldots, 15$). Results of physical, structural, thermal, and electrical measurements on the FVNBx glasses are presented.

2. Experimental

2.1. Glass preparation

The glasses of the desired composition were prepared by the conventional quick melt quenching technique. The starting materials were analytical grade reagents (AnalaR): H_3BO_3, Na_2CO_3, V_2O_5 and Fe_2O_3. The desired amounts of components were calculated according to the molar formula $x\mathrm{Fe_2O_3} + (30 - x)\mathrm{V_2O_5} + 30\mathrm{Na_2O} + 40\mathrm{B_2O_3}$ (abbreviated as FVNBx), where $x = 0$, 2, 3, 5, 7, 10, and 15 mol%. Appropriate amounts of Fe_2O_3, V_2O_5, Na_2CO_3, H_3BO_3 were well dry mixed and homogenized. The homogenized powders were melted in high purity alumina crucibles maintained at a temperature of 1050 °C in an electrically heated muffle furnace for about one hour. The alumina crucibles containing glass melts were swirled frequently to insure the homogeneity of the melt. The melts were then quickly quenched to lower temperature (≈ 300 °C) by pouring onto clean thick stainless steel (SS) plates preheated to ≈ 300 °C and subsequently pressing by another thick SS plate which was preheated to ≈ 300 °C to provide a better quenching rate, to reduce the strains during glass formation, and to obtain smooth surfaces of the plate-glass samples.

2.2. XRD

The as prepared samples were crushed to a fine powder and analysed using X-ray diffractometer (RigakuUltima IV) by employing Cu K_α radiations. The diffraction patterns were taken for two-theta (2Θ) from 20° to 80° at a scan speed of 2° per min.

2.3. Density measurement

The density D of the prepared glass samples was measured using the Archimedes principle at room temperature RT. The masses of the samples were measured by sensitive electronic balance (CAS CAVY 220). Xylene was used as an inert immersion liquid for measuring the buoyancy. The density was calculated using the formula

$$D = \frac{D_\chi W_a}{W_a - W_\chi}, \tag{1}$$

where D_χ, W_a, and W_χ are the xylene density, the weight of sample in air, and the weight of sample in buoyant liquid (here xylene), respectively. The molar volume V_m of each glass sample was calculated from data obtained for density. Overall, the formula $V_m = M/D$ was used, where M is molar mass of samples. The error in the density measurement was estimated as ± 0.01 g/cc (1 g/cc $= 10^3$ kg/m^3).

The theoretical optical basicity (Λ_{th}) was calculated by using the following relation:

$$\Lambda_{th} = X_{Fe_2O_3}\Lambda_{Fe_2O_3} + X_{V_2O_5}\Lambda_{V_2O_5} + X_{Na_2O}\Lambda_{Na_2O} + X_{B_2O_3}\Lambda_{B_2O_3}. \quad (2)$$

Here, X-terms are mole fractions and Λ-terms are optical basicity values of corresponding oxides, as assigned in literature [21–24].

2.4. Thermal analysis — differential scanning calorimetry

The differential scanning calorimetry (DSC) measurements of the samples in the bulk form were carried out in the temperature range of RT to 900 °C on a simultaneous thermal analyzer (Perkin Elmer STA 6000). The heating rate and nitrogen flow rate used to carry out the analysis were 10 °C/min and 100 ml/min, respectively. The differential temperature versus temperature (ΔT vs. T) and loss of mass versus temperature (Δm vs. T) were recorded and plotted. ΔT is taken as positive for the endothermic process. For loss of mass at different temperatures the mass of the sample under test is monitored as a function of temperature under the flow of hot dry nitrogen gas and then loss of mass in percent (Δm) is calculated as a function of temperature T.

2.5. FTIR

The FTIR spectra of the glass samples were recorded at room temperature on Perkin Elmer Frontier FTIR spectrometer in the mid-IR range, i.e., 400–4000 cm^{-1}. To record the infrared absorption spectra, KBr pellet technique was used. Base line and noise correction were achieved using Spectrum-10 software (provided with FTIR spectrophotometer). For peak identification of the processed spectra, multiple peaks fitting of each spectrum was carried out using *Origin-16 Pro* by adding peaks till the fit parameter became ≥ 0.999.

2.6. Electrical measurements

For these measurements, samples were cut and polished in the form of rectangular plates with the thickness ranging $\approx$ 1.0 mm. The broad faces of the regular shaped glass sample were painted by colloidal silver paste to remove any type of surface irregularity and to form two conducting electrodes. Since the samples showed high resistivity, the DC electrical conductivity of samples was measured by two probe method in the temperature range from RT to 200 °C with a heating rate of 1 °C/min. All the measurements were carried out on a Keithley low voltage source meter (Model 2401) in 4 wire mode and by applying a filter with repeated average of 50 points. A constant voltage of 20 V was applied across the high and low terminals while carrying out measurements. To avoid degradation of the samples, current was passed only for short periods (less than 20 s at a time). To avoid any sort of polarization effect, a toggle switch was used to reverse the polarity of applied voltage frequently at each measurement and the value of current was taken as an average of the maximal DC electric current values for both forward and reverse polarities. For a body with a constant cross-section A (m^2) and a constant thickness t (m) carrying a current I (ampere) under the applied DC voltage V (V), the volume conductivity σ (S/m) of the specimen can be determined from the elementary formula: $\sigma = [(tI)/(VA)]$. The estimated error in the measurement of σ is about 3%.

The measurements at frequency lower than 1 kHz has not been reported because of instrument limitations as it gives fully noisy and insignificant data for those frequencies. Impedance spectroscopy of samples was carried out on an impedance analyzer *(Hioki 3532-50 LCR HiTester)* in a frequency range of 1 kHz to 5 MHz and temperature range from RT to 200 °C with a heating rate of 1 °C/min controlled through a digital programmable electrical furnace. The experimental errors in the capacitance and resistance measurements were estimated to be ± 0.1 pF and ± 10 Ω, respectively. The complex dielectric permittivity, $\varepsilon^* = \varepsilon' - j\,\varepsilon''$ (with ε' and ε'' as the real and the imaginary part of dielectric constant), dielectric loss tangent $\tan(\delta)$ and AC conductivity σ_{ac} were calculated from the standard relations

$$\varepsilon' = Ct/\varepsilon_0 A, \quad (3a)$$

$$\varepsilon'' = \varepsilon' \tan(\delta), \quad (3b)$$

$$\sigma_{ac} = 2\pi f \varepsilon' \varepsilon_0 \tan(\delta), \quad (3c)$$

where the symbols used are capacitance C (in F), area of broad cross-section A (in m^2), thickness t (in m), permittivity of free space $\varepsilon_0 = 8.854$ pF/m, and f (in Hz) is the frequency of AC voltage used in the experiment.

3. Results and discussion

3.1. XRD

The glass samples were tested by XRD to confirm the glassy nature of all the samples. The XRD patterns are shown in Fig. 1. All the samples have shown only characteristic glass XRD patterns. No crystalline phase is detected by XRD. This confirms the non-crystalline nature of the samples prepared for study.

3.2. Density

Measured values of density and calculated values of molar volume are collated in Table I. There is a distinct variation in the densities of the glasses depending on the iron and vanadium contents. Figure 2 displays the variation of the density as a function of iron content x.

TABLE I

Composition, density D, molar volume V_m, theoretical optical basicity Λ_{th} of FVNBx ($x = 0, 2, 3, 5, 7, 10, 15$) glasses.

x	Sample	Molar weight [g]	Density [g/cm^3]	Molar volume [cm^3/mol]	Optical basicity
0	0 Fe_2O_3 + 30 V_2O_5 + 30 Na_2O + 40 B_2O_3	136.78	2.42	56.43	0.902
2	2 Fe_2O_3 +28 V_2O_5 + 30 Na_2O + 40 B_2O_3	136.47	2.53	53.87	0.9016
3	3 Fe_2O_3 + 27 V_2O_5 + 30 Na_2O + 40 B_2O_3	136.32	2.57	53.12	0.9014
5	5 Fe_2O_3 +25 V_2O_5 + 30 Na_2O + 40 B_2O_3	136.00	2.59	52.41	0.901
7	7 Fe_2O_3 + 23 V_2O_5 + 30 Na_2O+ 40 B_2O_3	135.69	2.63	51.55	0.9006
10	10 Fe_2O_3 + 20 V_2O_5 + 30 Na_2O+ 40 B_2O_3	135.22	2.67	50.59	0.9000
15	15 Fe_2O_3 + 15 V_2O_5 + 30 Na_2O + 40 B_2O_3	134.44	2.72	49.46	0.899

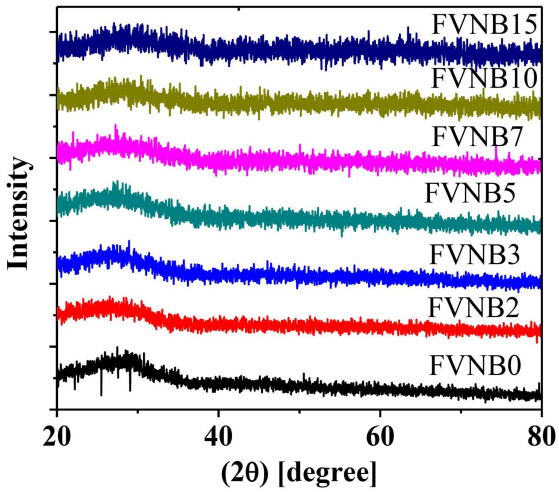


Fig. 1. XRD spectra of FVNBx ($x = 0$, 2, 3, 5, 7, 10, 15) glasses recorded at RT.

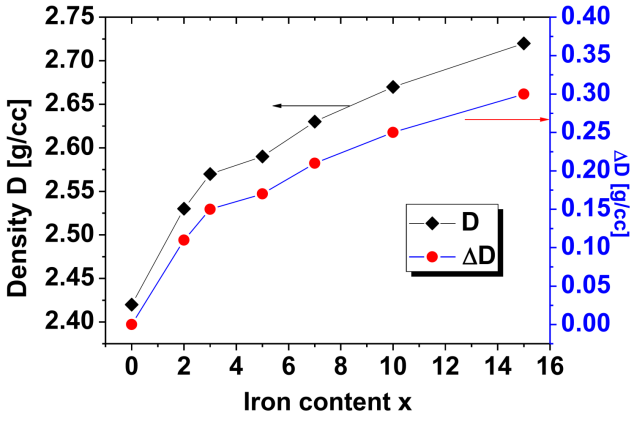


Fig. 2. Density variation of FVNBx ($x = 0$, 2, 3, 5, 7, 10, 15) glasses. ΔD represents the change in density with respect to FVNB0.

As can be seen, the overall change in density is around 0.3 g/cc when replacing V_2O_5 with increasing concentration of Fe_2O_3 (from $x = 0$ to $x = 15$). The increase in density and decrease in molar volume with the increase in iron content is justified because glass modifier V is being replaced by heavier metal iron and a better filling factor is created by Fe–O coordination in the glass microstructure by acting both as glass former and glass modifier.

The optical basicity proposed by Duffy and Ingram was used as a measure of acid-base properties of oxide glasses and is expressed in terms of the electron density carried by oxygen [25–28]. The values of theoretical optical basicity for the prepared compositions (as shown in Table I) indicate a nominal decrease with an increase in Fe_2O_3 content at the cost of V_2O_5 in the glass. The decreasing optical basicity by addition of Fe_2O_3 in place of V_2O_5 indicates a marginal decrease in the ability of oxygen to donate negative charge to the bonding. It points towards an increase in covalent character of the bonds in the glass network with increase in Fe_2O_3 content. This property may affect the electrical conduction of the glasses and a decrease in conductivity is anticipated.

3.3. Thermal analysis

From the ΔT versus T scans the glass transition T_g could be estimated from the broad endothermic shift as shown in Fig. 3 for all FVNBx ($x = 0$, 2, 3, 5, 7, 10, 15) glasses. The glass transition temperature T_g is considered to be a characteristic of structural relaxations taking place in the glass network and depends upon many factors, namely, bond strength, cross linking density, enthalpy of atomization, nature of the structural units constituting the glass network and their connectivity way [29, 30]. It was observed that the glass transition lies approximately between 500 °C and 550 °C (±25 °C) for all the glasses prepared. Further, by close analysis of Fig. 3, an increase in T_g is indicative with the increase of iron content x. This may be due to the reason that with increase in iron content, the inter atomic distance decreases and overall structure become more closely packed, i.e., density and bond strength increased and more energy is required to dissociate Fe–O and V–O bonds. This will cause an increase in transition temperature T_g with increase in iron content [31, 32]. This indicates that the Fe_2O_3 acts both as glass former and as glass modifier. To determine the thermal stability of the prepared glasses a known amount of the sample was heated with dry

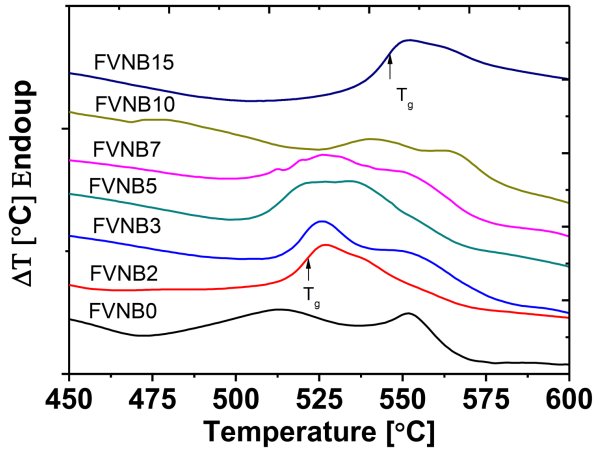


Fig. 3. Differential temperature ΔT vs. temperature T scans of FVNBx (x = 0, 2, 3, 5, 7, 10, 15) glasses. ΔT upward indicates the endothermic process.

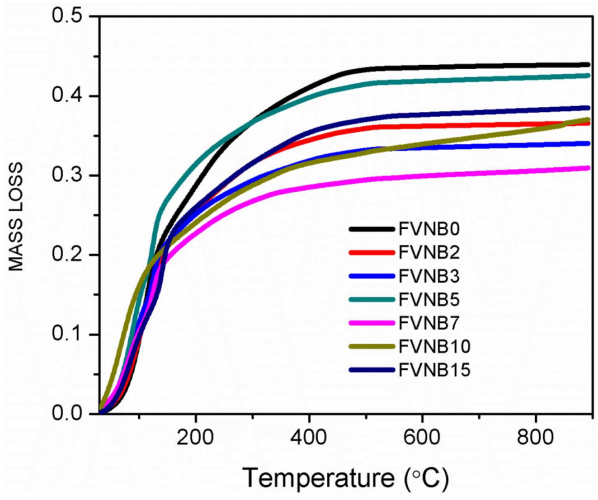


Fig. 4. Percent mass loss vs. T scans of FVNBx (x = 0, 2, 3, 5, 7, 10, 15).

nitrogen gas flow at a constant rate and the corresponding mass loss percent $\Delta m = 100(m_T - m_{RT})/m_{RT}$ was measured at different temperatures. Here, m_{RT} and m_T are the masses of the sample at RT and at temperature T, respectively. The measurements were performed on powdered samples. As prepared glass wafers were crushed to fine powders in a mortar and pestle under ambient conditions which were subsequently used for the mass loss measurements. The observed mass losses (in %) vs. temperature are shown in Fig. 4.

The observed percent mass losses lie in the range from 0.25 to 0.45 in the glass samples under investigation in the temperature range of study. The small mass loss is attributed to the evolution of the adsorbed moisture and ambient gases during the handling of the glass powders. It may also be mentioned that at RT no detectable mass loss was observed even after one year of the vacuum dry storage. This confirms that the prepared glasses are chemically stable over the temperature range of study.

3.4. FTIR study

The FTIR spectra of FVNBx glass systems at RT in the frequency range of 400–1600 cm^{-1} are shown in Fig. 5.

According to an assumption in vibrational spectroscopy of the solid state, the vibrations of a specific group of atoms in a lattice are regarded as relatively independent of motions of the rest of the atoms. In FTIR specific groups of atoms are characterized by distinct normal modes of vibrations and are displayed as absorption bands. The FTIR spectra show multiple absorption peaks which are attributed to different vibrational modes of various multi-atomic species existing the glass matrix (see Table II).

To analyze the FTIR spectra more accurately, multiple peak fittings of the FTIR spectra have been done using Origin 8.1Pro. Peaks were fitted to the simulated curve till the fit parameter became > 0.999. Gaussian peak fitted/convoluted spectra for all the samples are shown in Fig. 6. Through the peak fitting, the peak position/centre (cm^{-1}) and the relative area (intensity) of the peaks were determined. The data thus obtained is collated in Table III.

The FTIR spectra of sodium–iron–boro–vanadate glasses are expected to mainly consist of three absorption regions: (i) high frequency range 1200–1600 cm^{-1} attributed to the vibrations of B–O in borate units, in which boron is connected to three oxygen atoms (both bridging (BO) and non-bridging (NBO) types), (ii) mid frequency range 800–1200 cm^{-1}, which are mainly formed by tetrahedral coordinated boron (BO_4), and (iii) low frequency ranges 400–800 cm^{-1}, which are due to B–O–B and V–O–V bending and stretching of Fe–O (tetrahedral FeO_4 and octahedral FeO_6 units).

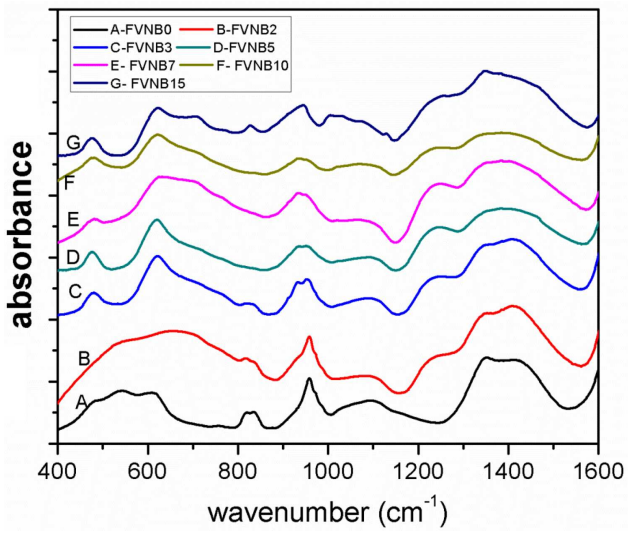


Fig. 5. FTIR spectra of FVNBx (x = 0, 2, 3, 5, 7, 10, 15) glasses recorded at RT.

TABLE II

Expected assignment of IR bands in the spectra of FVNBx [25, 26].

Peak position [cm^{-1}]	Assignment
∼ 1200–1400	stretching of B–O of trigonal boron (BO_3)
∼ 1250	stretching of trigonal boron connected to V atoms in borovanadate species $(B_2V_2O_9)^{2-}$
∼ 900–1100	stretching B–O of tetrahedral boron (BO_4)
∼ 950–1000 (sharp)	V=O stretching of vanadyl $(VO)^{2+}$
∼ 900–1050	stretching of four connected five-coordinated vanadium species (V^{4-}) $(VOO_4)^{2-}$
∼ 790–800	Stretching of two connected four coordinated vanadium species (V_2^-) $(VOO_2O)^-$
∼ 700–710	vibrations involving VOO_3^-
∼ 660	stretching modes of Fe–O of tetrahedral FeO_4
∼ 610–700	B–O–B bending of trigonal boron atoms
∼ 460–530	stretching modes of Fe–O of octahedral FeO_6
∼ 430–450	V–O–V bending modes

The absence of the prominent peak ≈ 800 cm^{-1} in the IR spectra of the present glasses is an indication of the missing six-member boroxol rings –(B_3O_3)– in the glass network. Based on the relative intensity the observed absorption bands are categorized as weak (W), medium (M) or strong (S) (see Table III).

In the present glasses, the main IR absorption peaks are present around: {614(M), 650–702(S), 921FVNB15, 944–960, 1050–1096, 1200–1275, 1330–1400, 1445, 1474, 1540, 1592} cm^{-1}. Asymmetric vibrations of B–O bonds from pyro- and ortho-borate groups in BO_3 trigonal units are confirmed from the peaks in the range 1200–1500 cm^{-1}. The IR spectrum of vitreous V_2O_5 is found similar to that of the crystal. It shows a sharp band around 950–1000 cm^{-1}, which means that the isolated V=O bonds from the VO_5 trigonal bipyramids are preserved also in the vitreous structure. The strong bands around 650–702 cm^{-1} are attributed to the bending vibrations of the O–V–O bond. The bands around 650–702 cm^{-1} are assigned to the combination vibrations of $(VO_3)_n$ single chain or due to antisymmetric vibrational modes of V–O–V group. The bands around 929–1007 cm^{-1} are attributed to the symmetric stretching vibrations of V–O–V. The main coordination polyhedron for vanadium in the crystal structures is the VO_4 tetrahedron with three terminal V–O bonds forming VO_3 groups. In the vitreous V_2O_5 counterpart, there is a band at ≈ 920 cm^{-1}, which is close to the symmetrical stretching vibrations of VO_3 unit, in the spectra of crystals. This band is observed in the FTIR spectrum of FVNB15 glass. The bands around 950–970 cm^{-1} are attributed

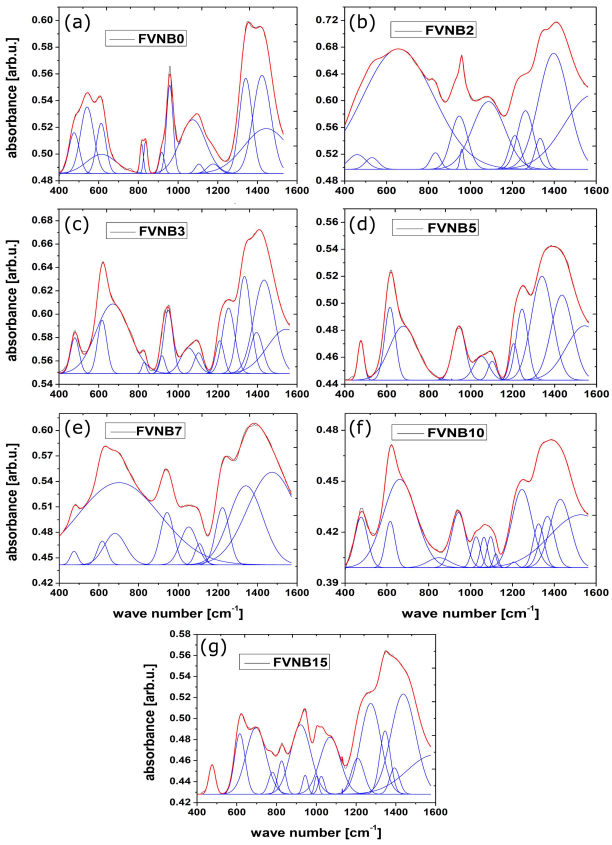


Fig. 6. Gaussian deconvoluted FTIR spectra for (a) FVNB0, (b) FVNB2, (c) FVNB3, (d) FVNB5, (e) FVNB7, (f) FVNB10, (g) FVNB15 glasses.

not only to the VO_5 groups but also to the branched VO_4 groups having one $(V=O)^{2+}$ double bond. This band has a medium to weak intensity in the present glasses. The weak presence of ν (V=O) band indicates that majority of the V exists in V^{5+} and V^{4+} valence is in minority. Such observation was confirmed by Saetova et al. in Li_2O–V_2O_5–B_2O_3 glasses. The absorption band at 1074 cm^{-1} in iron free FVNB0 glass shifts to a higher frequency (1086 cm^{-1}) in FVNB2 due to incorporation of two mole percent of Fe_2O_3. A shift to a higher frequency is attributed to shortening of B–O bond in di- and penta-borate linkages. However, this band remains stable 1152 cm^{-1} in FVNB3, FVNB5, and FVNB7 which indicates that a weakening of B–O bond has taken place for these concentrations of Fe_2O_3. For higher concentration of Fe_2O_3 ($x = 10$ and $x = 15$) a gradual shift towards higher frequency is observed which indicates a trend of strengthening of B–O bond for glasses containing higher iron content. The strong high frequency bands between 1300–1600 cm^{-1} attributed to O–B–O bonds of tetrahedral BO_4 and trigonal BO_3 bonds also show significant changes due to change in Fe_2O_3/V_2O_5 composition. The irregular trends in shifts of the frequencies confirm the changes in the structure of the glass network mainly through strengthening and/or weakening of the oxygen bonds (bridging and/or non-bridging).

TABLE III

FTIR peak centre [cm^{-1}] and relative peak area (intensity) for FVNBx glasses. The relative intensity is shown in parenthesis along with the peak centre [cm^{-1}]. Peaks with intensity (1)–(4) are termed as weak, with (5)–(9) as medium and above (9) as strong.

FVNB0	FVNB2	FVNB3	FVNB5	FVNB7	FVNB10	FVNB15
476(4)	460(1)	479(3)	476(2)	475(1)	477(5)	476(1)
542(9)	530(1)	–	–	–	–	–
613(5)	–	–	–	–	–	–
614(6)	–	617(5)	615(3)	617(1)	616(3)	615(5)
–	656(39)	671(20)	679(13)	681(4)	661(22)	697(10)
–	–	–	–	702(39)	–	–
817(1)	–	–	–	–	–	780(1)
837(1)	833(1)	829(1)	–	–	848(2)	826(2)
918(1)	–	919(1)	910(1)	–	–	921(10)
–	947(4)	–	944(6)	944(4)	942(6)	944(1)
958(1)	959(1)	951(5)	–	–	–	–
–	–	–	–	–	–	1001(1)
–	–	1020(1)	–	–	1026(2)	1025(1)
–	–	1052(3)	1051(3)	1053(4)	1064(2)	1069(9)
1074(14)	1086(9)	–	–	–	1096(2)	–
1105(1)	–	1104(2)	1102(1)	1107(1)	1120(1)	1130(1)
1177(1)	–	–	–	–	–	–
–	1212(2)	1210(2)	1205(3)	1224(4)	1207(1)	1208(3)
–	1263(4)	1255(3)	1245(4)	–	1245(11)	1274(14)
1342(12)	1333(1)	1335(11)	1340(20)	1341(15)	1325(4)	1346(5)
–	1397(16)	1395(5)	–	–	1367(5)	1393(2)
1423(18)	–	–	–	–	1429(11)	–
–	–	1434(18)	1435(17)	–	–	1438(19)
1445(21)	–	–	–	1474(26)	–	–
–	1584(19)	1546(15)	1543(18)	–	1527(25)	1592(16)

Another interesting feature is displayed by the shifts in the low frequency bands. The negligible shift in bands assigned to V–O bonds indicates a stable nature of the V–O bonds in the glass network of the glasses under study. However, a band ≈ 655 cm^{-1} assigned to Fe–O bond of tetrahedral FeO_4 units reflects a regular increase in frequency with x (except for $x = 10$). This confirms that the shortening of Fe–O bond is occurring in the tetrahedral FeO_4 units of the glass network as the Fe_2O_3 content is increased. Therefore the analysis of the FTIR spectra of the samples revealed significant changes in the features of the absorption bands which are a confirmed signature of the micro-structural changes in the glass network due to change of Fe/V composition. Therefore in the present glass system Fe_2O_3 acts both as glass modifier as well as glass former.

3.5. Electrical measurements

The temperature dependence of DC electrical conductivity σ of as prepared glasses are presented in Fig. 7. This dependence of the DC conductivity obeys the well known Arrhenius formula [16, 27, 28], namely

$$\sigma = \sigma_0 \exp\left(-\frac{W}{k_B T}\right), \tag{4}$$

where W is the activation energy for conduction, σ_0 is a constant for given glass, k_B is the Boltzmann constant and T is the temperature in kelvin (K). The values of W and pre-exponential factor σ_0, were evaluated by the least squares fitting of the experimental data, using the relation

$$\log(\sigma) = \log(\sigma_0) - \frac{W}{2303 k_B}\frac{1000}{T}. \tag{5}$$

This linear relationship between conductivity and reciprocal of temperature indicates that the activation energy is independent of temperature in the range studied here. The fact that σ increases with increase in temperature establishes the semiconducting type of behaviour for electrical conductivity. The calculated values of σ (at 80 °C and 180 °C), σ_0 and W are displayed in Table IV. The x dependence of DC conductivity at 80 °C and 180 °C are shown in Fig. 8 for comparison. The activation energy W vs. x plot is shown in Fig. 9. At 80 °C the conductivity is found to decrease as x is increased. The activation energy however, shows a reverse correlation with x, which is consistent with small polaron hopping (SPH) model [27, 28, 33] and explains the semiconducting nature of the glasses under study. At 180 °C the conductivity of FVNB10 glass increases as compared

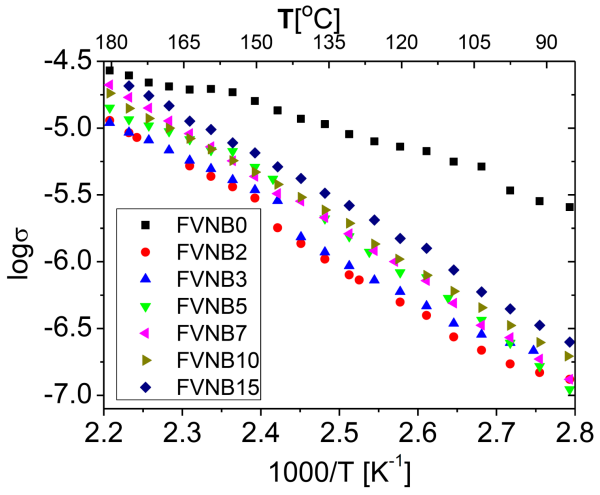


Fig. 7. Variation of $\log(\sigma_{dc})$ [S/m] as a function of reciprocal of temperature for FVNBx (x = 0, 2, 3, 5, 7, 10, 15) glasses.

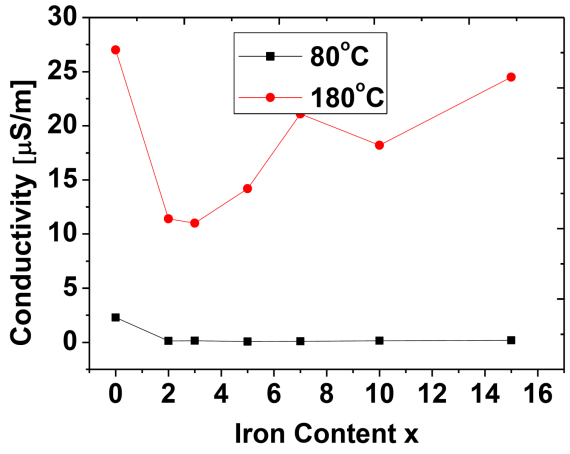


Fig. 8. DC conductivity at 80 °C and 180 °C for the FVNBx (x = 0, 2, 3, 5, 7, 10, 15) glasses.

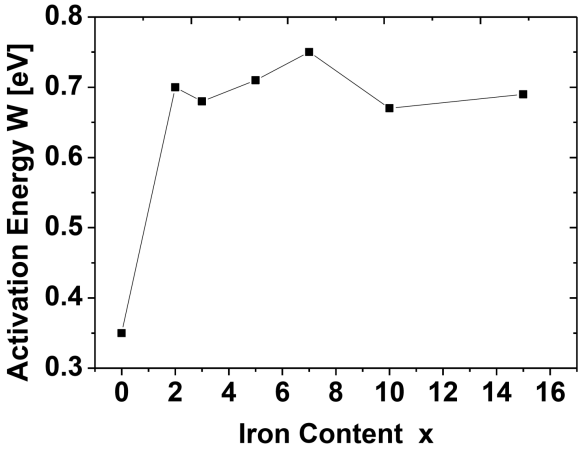


Fig. 9. Activation energy W of the FVNBx (x = 0, 2, 3, 5, 7, 10, 15) glasses.

TABLE IV

DC conductivity σ, prefactor σ_0 and activation energy W for FVNBx (x = 0, 2, 3, 5, 7, 10, 15) glasses

Sample	σ at 80 °C [μS/m]	σ at 180 °C [μS/m]	σ_0 [S/m]	W [eV]
FVNB0	2.3	27.0	0.21	0.35
FVNB2	0.12	11.4	744.6	0.70
FVNB3	0.14	11.0	397.2	0.68
FVNB5	0.07	14.2	1490.7	0.71
FVNB7	0.09	21.1	4243.3	0.75
FVNB10	0.01	18.2	564.9	0.67
FVNB15	0.02	24.5	1100.3	0.69

to FVNB7, but it decreases continually with further increase in x up to 15. The linear relationship between conductivity and reciprocal of temperature indicates that the activation energy is independent of temperature. As compared to the iron free glass FVNB0 the conductivity of iron containing $x > 0$ glasses decreases drastically (by an order of magnitude) with the inclusion of Fe_2O_3 in the glass network (the average value of σ at 80 °C being is 0.08 μS/m). The prominent conduction mechanism in V_2O_5 is due to electron hopping between V^{3+}, V^{4+}, and V^{5+} states and in mechanism in Fe_2O_3 is due to hopping in Fe^{2+}, Fe^{3+} ions. Recently, Pietrzak et al. [34] had studied the Li_2O–FeO–V_2O_5–P_2O_5 system by the Mössbauer spectroscopy and thermoelectric power measurements to find the most prominent hopping centre between iron or vanadium ions, which one is dominating the charge carrier transport in glass system. It has been found that Fe^{2+}/Fe^{3+} ratio does not have major influence on electronic conductivity possessed by dominating V^{4+}/V^{5+} hopping sites. Similar phenomenon is observed in conduction processes of presently prepared glass systems. With the increase in Fe_2O_3 content and consequent decrease in V_2O_5 content, the decrease in electrical conductivity implies that V^{4+}/V^{5+} ions in the present glass system create structural changes such that the charge carrier transport responsible for electrical conduction in the glass network is hindered. This may be translated to the increase in small polaron hopping (SPH) range/distance and resistance to polaron orientation motion (both jump as well as frequency). Within the iron containing glasses $x > 0$, the conductivity is minimal for FVNB10 glass at 80 °C and for FVNB3 at 180 °C, whereas it is maximal for FVNB3 at 180 °C and for FVNB15 at 180 °C. The total conductivity is due to carrier hopping in both types of ions, i.e., iron (less dominating) and vanadium (more dominating). Here, iron ions is replaced by more conducting vanadium ions, which results into the decreasing trend of electrical conductivity values [35].

Also the reduction in vanadium content may cause gradual reduction of conductivity if conduction due to electronic hopping $V^{4+} \Longleftrightarrow V^{5+}$ is envisaged due to expected increase in V–V distance with x. But the abrupt decrease in conduction upon replacing V by small amount

of Fe does not support this mechanism, too. Therefore, the conduction appears to be through the charge generated on the non bonding oxygens (NBOs) in the glass network through small range polaron hopping (SRPH) mechanism. The average activation energy for the iron containing glasses (i.e., $x > 0$) is 0.7 eV which may be assumed to be same for all the iron containing glasses FVNBx ($x = 2, \ldots, 15$) yet the conductivity of FVNB15 at 180 °C appears to be approaching the value of conductivity of iron free FVNB0 glass. This observation can be explained by the assumption that the small range polaron hopping (SRPH) frequency increases in the iron containing glasses as the temperature is raised. An increase in SRPH frequency at higher temperatures is justified due to available thermal excitation.

3.6. AC conductivity

Figure 10a shows the frequency dependence of AC conductivity σ_{ac} for FVNB0, FVNB5, and FVNB15 compositions at $T = 200$ °C and Fig. 10b shows the frequency dependence of σ_{ac} for FVNB5 at different temperatures. These are taken as representative cases since other samples also show similar trends. At lower frequency, conductivity is low and is almost independent of frequency. This is explained on the basis of space charge polarisation effect which is responsible for the large bulk resistance at lower frequencies. As a consequence, a plateau region is obtained (see Fig. 10), which shows the frequency independent part of conductivity or the absolute value of impedance. With further increase in frequency, dispersion at higher frequencies (in MHz range) has been observed at all the temperatures and in all the samples studied in this work. The dispersion at higher frequency shifts towards higher frequencies with an increase in temperature. Further, the conductivity spectrum has found [36] to obey the Jonscher power law (JPL) of the AC conductivity in glasses, which is mathematically written as

$$\sigma = \sigma_0 + A_1(f)^{s1} + A_2(f)^{s2}. \tag{6}$$

Here, σ_0 is the conductivity at limiting zero frequency (termed as frequency independent DC component of conductivity), A_1 and A_2 are constants, and s_1 and s_2 are the characteristic temperature dependent parameters. The power law exponents s_1 and s_2 (lies in the range

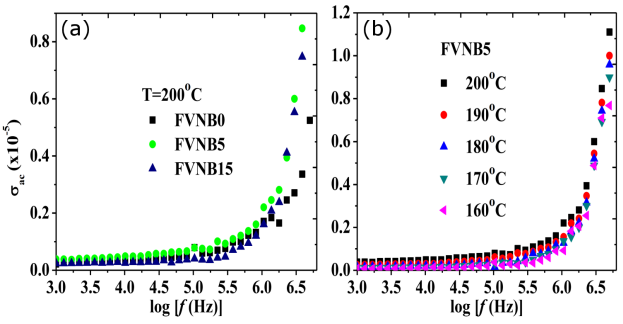


Fig. 10. Variation of σ_{ac} vs. $\log(f)$ [Hz] for (a) compositions FVNB0, FVNB5 and FVNB15 at temperature 200 °C, (b) FVNB5 at different temperatures.

TABLE V

Dielectric parameters and AC electrical conductivity of FVNBx glasses at 200 °C and at 5 MHz.

Sample	ε'	ε'' ($\times 10^{-2}$)	$\tan(\delta)$	M'	M''	σ_{ac} [μS/m] (at 5 MHz)
FVNB0	0.17	0.19	0.11	5.920	0.67	5.25
FVNB2	0.05	0.12	0.23	17.50	3.99	3.45
FVNB3	0.05	0.06	0.12	19.13	2.27	1.70
FVNB5	0.03	0.06	0.16	28.67	4.67	1.54
FVNB7	0.07	0.03	0.04	13.69	0.61	0.91
FVNB10	0.13	0.04	0.03	7.50	0.21	1.03
FVNB15	0.13	3.30	0.26	7.33	1.88	9.15

$0 < s_1, s_2 < 1$) represent the degree of interaction between the mobile ions [37]. The AC conductivity values at $T = 200$ °C and frequency 5 MHz for all prepared compositions are reported in Table V. The observed trend of change of σ_{ac} values is due to hindrance offered by mixed effect of iron and vanadium ions which further results into the polaronic conduction in prepared glasses [36]. It can be observed that σ_{ac} increases with increase in temperature, which is possibly due to depletion of space charge effect and thermally activated hopping of charge carriers [38–40]. It is quite interesting to note that the value of frequency independent conductivity measured at 1 kHz at 200 °C for iron containing $x > 0$ glasses is ≈ 0.07 μS/m, which is very near to the average DC conductivity (≈ 0.08 μS/m) calculated for prepared glasses at 80 °C (as shown in Fig. 8).

4. Dielectric properties

4.1. Dielectric loss

The variation of dielectric constant (ε') as a function of frequency at $T = 200$ °C is illustrated in Fig. 11a. For all the prepared compositions variation of ε' with frequency is not significant except in case of FVNB5 (shown in Fig. 11b). The frequency dependence of dielectric loss ε'' is shown in Fig. 12a, as a function of composition x, while in Fig. 12b the temperature dependence of ε'' for FVNB5 is shown. Various polarization mechanisms in dielectrics are: (i) electronic, (ii) ionic, (iii) dipole orientation/deformation in polar dielectrics, and (iv) migrational/space charge polarization [27]. The contribution of electronic polarization is negligible as compared to the other contributions due to ionic polarization and in non-polar dielectrics the main contribution comes from the last mechanism. The ionic polarization is frequency independent up to far IR frequencies. However, the orientation/deformation polarization and the migrational/space charge polarization are frequency dependent and usually decrease at higher frequencies. The ionic polarization is temperature independent but the orientational and the migrational polarizations are temperature dependent and while the former is expected to decrease due to thermal agitation the later is expected

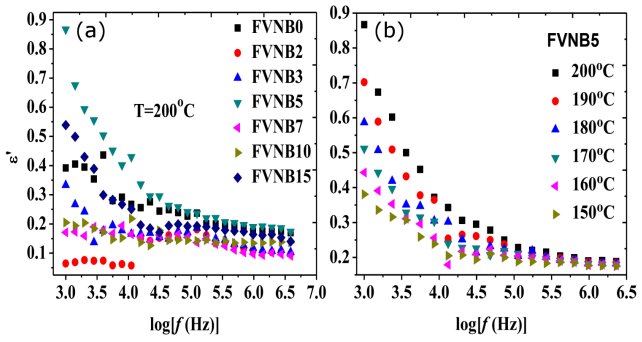


Fig. 11. Variation of ε' vs. $\log(f)$ [Hz]: (a) different compositions at 200 °C, (b) FVNB5 at different temperatures.

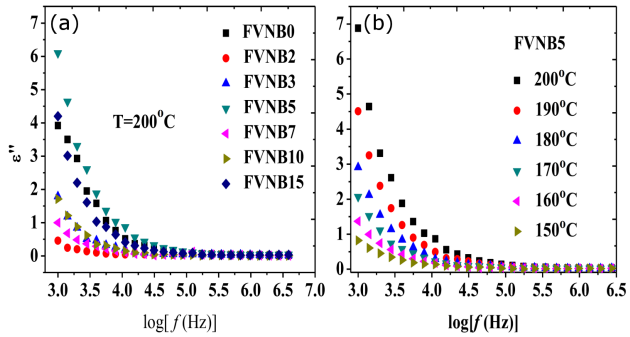


Fig. 12. Variation of ε'' vs. $\log(f)$ [Hz]: (a) different compositions at temperature 200 °C, (b) FVNB5 at different temperatures.

to increase with the rise in temperature. Study of literature [17, 18, 27, 28, 33, 38–40] suggests that in dielectric permittivity ε^* of a sample is due to (i) rotation of ions around their negative sites and (ii) short range dipole transport (hopping).

Analysis of ε' and ε'' shows that both the parameters have higher values at lower frequencies and the values become very small (frequency independent) at higher frequencies which is a signature of migration polarization in non-polar dielectrics. At lower frequencies, the space charge accumulation occurs at the electrolyte (glass)–electrode interface, which does not permit further transfer of the ions through the dielectric. At the higher frequencies, the periodic reversal of electric field occurs and the polarization due to accumulation of charge decreases and causes a decrease in the values of ε' and ε'' with the increase in frequency [37–41]. The variations of ε' and ε'' as a function of frequency for FVNB5 at different temperatures are shown in Fig. 11b and Fig. 12b, respectively. The effect of temperature change is significantly observed at lower frequencies on ε' and ε''. At lower frequencies, the value of ε' and ε'' increases with temperature whereas at higher frequencies, variation in ε' and ε'' has been found more or less independent of temperature and frequency. This occurs due to weakening of bonds

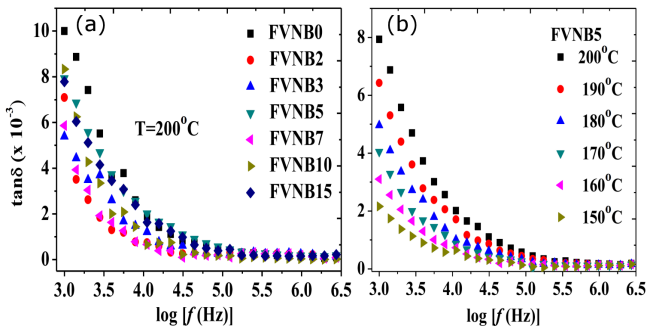


Fig. 13. Variation of $\tan(\delta)$ vs. $\log(f)$ [Hz]: (a) different compositions at 200 °C, (b) FVNB5 at different temperatures.

and an increase in thermal agitation that causes a disturbance in the orientational vibrations of the molecular groups in glass network. Therefore, space charge polarization decreases with increase in temperature [39].

4.2. Loss tangent ($\tan(\delta)$)

Figure 13a shows the frequency dependence of dielectric loss tangent ($\tan(\delta)$) for all the prepared FVNBx compositions. The variation of $\tan(\delta)$ with frequency at constant temperature exhibits similar trends as for ε'' which could be due to the fact that as the frequency increases, the polarizability contribution from ionic and orientation sources decreases and finally disappears due to their inertia. Figure 13b shows the frequency dependence of dielectric loss tangent for FVNB5 at different temperatures. The dielectric loss is seen to decrease for $x = 7$ and $x = 10$ as compared to other glasses. The variation of dielectric parameters upon changing the Fe/V content verifies the structural changes in the glass networks although no regular trend is visualized.

4.3. Modulus formalism

To analyze the relaxation data, we often need to represent the dielectric as a combination of pure resistances and capacitances and the manner of their connection is assumed. Relaxations in such circuits can proceed through two independent processes known as serial and parallel processes. Macedo et al. [40] proposed on the basis that impedances are additive in a series process and used electrical modulus, $M^* = 1/\varepsilon^*$, as appropriate parameter for the analysis of conductivity relaxation. Therefore, M^* is also resolved into real and imaginary parts as

$$M^* = \frac{1}{\varepsilon^*} = \frac{\varepsilon'}{(\varepsilon')^2 + (\varepsilon'')^2} + \frac{\varepsilon''}{(\varepsilon')^2 + (\varepsilon'')^2}. \quad (7)$$

Redefining Eq. (3) by using ε_s to represent frequency independent dielectric constant and introducing $\tau_\sigma = \varepsilon_0\varepsilon_s/\sigma_0$ and $M_s = 1/\varepsilon_s$, we come to the relation $M^* = M' + M''$, written as

$$M^* = M_s\left(\frac{\mathrm{i}\omega\tau_\sigma}{1 + \mathrm{i}\omega\tau_\sigma}\right) = M_s\left(\frac{(\omega\tau_\sigma)^2 + \mathrm{i}\omega\tau_\sigma}{1 + (\omega\,\tau_\sigma)^2}\right). \quad (8)$$

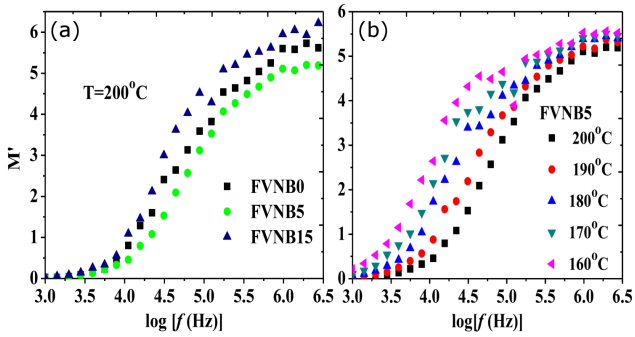


Fig. 14. Variation of M' vs. $\log(f)$ [Hz] for (a) compositions FVNB0, FVNB5 and FVNB15 at temperature 200 °C, (b) FVNB5 at different temperatures.

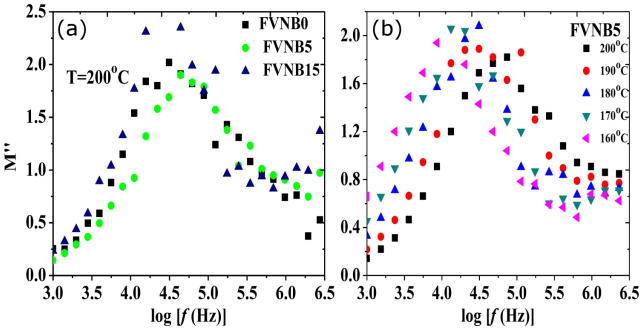


Fig. 15. Variation of M'' vs. $\log(f)$ [Hz]: for (a) compositions FVNB0, FVNB5 and FVNB15 at 200 °C, (b) FVNB5 at different temperatures.

In Eq. (8), τ_σ is known as conductivity relaxation time and is used to represent electric field decay in an ionic glass as $E = E_0 \exp(-t/\tau_\sigma)$. It suggests that in the limits $\omega\tau_\sigma \ll 1$ and $\omega\tau_\sigma \gg 1$, M' becomes zero and $1/\varepsilon_s$, respectively and so as M''. But at $\omega\tau_\sigma = 1$, M'' gives rise to a peak which is called a relaxation (resonance) peak.

The variation of real parts of the electric modulus M' and imaginary parts of the electric modulus M'' as a function of frequency for FVNB0, FVNB5, and FVNB15 at 200 °C are shown in Fig. 14a and Fig. 15a, respectively. The variation of M' and of M'' as a function of frequency for sample FVNB5 at various temperatures are shown in Fig. 14b and Fig. 15b, respectively. The electric modulus M' profile in Fig. 14 shows that the curve almost tends to zero in the low frequency range and it is due to the suppression of electrode polarization. At higher frequencies, in turn, it reaches M'_∞ because of relaxation process over a range of frequencies. Also the variation of M' shows that M' becomes almost constant at higher frequencies. The zero value of M' at low frequencies marks the non-existence of restoring force for the flow of charge carriers. The shape of M'' curve in Fig. 15 is broader than ideal Debye curves and asymmetric in nature. The maximum in the imaginary part of the electric modulus M'' shifts to higher frequencies with

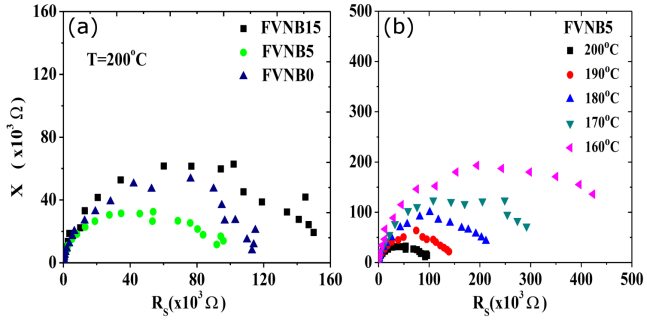


Fig. 16. Nyquist diagrams vs. R_s for (a) compositions FVNB0, FVNB5 and FVNB15 at 200 °C, (b) FVNB5 at different temperatures.

increase in temperature, which shows that the number of charge carriers increases by thermal activation achieved through phonon absorption. The peak frequency corresponding to $M''_{\max}$ corresponds to relaxation time τ_σ, which satisfies the Arrhenius law. The smaller value of M'' at low frequencies indicates negligible contribution of electrode polarization to the electric modulus. Two apparent relaxation regions appear above and below $M''_{\max}$, the low frequency region ($f < f_m$) determines the range in which charge carriers are mobile over long range distance, whereas in high frequency region ($f > f_m$), the charge carriers are confined to potential well, being mobile on short distances and associated with the relaxation polarization process [27, 42–44]. An effective interaction between charge carriers establish correlation between ions and governs non-bridging oxygen sites in the glass. This is possibly the prime cause of appearance of the non-Debye conductivity relaxation in glass compositions under study.

4.4. Impedance spectroscopy

The Nyquist plots [25] for FVNB0, FVNB5, and FVNB15 at 200 °C are shown in Fig. 16a. The impedance data analysis shows that a single semicircle arc is present in all the prepared compositions and there is no residual semicircle at lower frequency, which attributes to the single conduction mechanism [43, 44]. The centre of each semicircle is observed to be depressed below R_s axis which indicates that the associated relaxation of ions is non-Debye in nature [19, 37]. From analysis, the semicircle arcs passing through the origin of impedance plots and the intercept at lower frequency on the R_s axis give bulk resistance R_b. The value of R_b is highest for FVNB0 (i.e., 0 mol% of Fe_2O_3). With increase in concentration of Fe_2O_3, the value of R_b first decreases and then increases, i.e., a non-linear variation of R_b is observed, which indicates the mixed effect of iron and vanadium in the glassy matrix. The Nyquist plots for FVNB5 at different temperatures are shown in Fig. 16b. A single semi-circle in these plots indicates a single conduction mechanism in the FVNB5 sample. The depression of the semi-circle in these plots on the real axis shows that the relaxation

mechanism is non-Debye in nature. The intercepts on the real axis corresponding to zero value of the imaginary axis in the Nyquist plot gives us the bulk resistance of the sample excluding the polarization effects [27, 44]. The values of bulk resistance were found to decrease with increase in temperature and hence DC conductivity of the sample increases with increase in temperature depicting the semi-conducting nature of the sample under study.

5. Conclusion

The structural, physical, and electrical properties of $x\mathrm{Fe_2O_3} + (30 - x)\mathrm{V_2O_5} + 30\mathrm{Na_2O} + 40\mathrm{B_2O_3}$ (abbreviated as FVNBx), where $x = 0, 2, 3, 5, 7, 10$, and 15, have been studied by XRD, density measurement, FTIR, dielectric, and AC and DC electrical conductivity measurements. Density D of the glass samples increases and molar volume V_m decreases with the replacement of $\mathrm{V_2O_5}$ by $\mathrm{Fe_2O_3}$ content by mole fraction x. This is attributed to the decrease in NBOs due to the conversion of trigonal $\mathrm{BO_{3-}}$ structural units into $\mathrm{BO_{4-}}$ tetrahedral and higher oxygen coordination units (pyro- and orthoborates/vanadates) indicating the increase in compactness of the glass network structure. The marginal decrease in the theoretical optical basicity Λ_{th} is indicative of the decrease in ability of oxygen to donate electron. No boroxol rings formation were observed in the structure of these glasses. This suggests that glasses under study consist of randomly connected M–B–O groups, where M=O/Fe/V/Na.

The FTIR spectra and thermal analysis of the present glass systems indicate that $\mathrm{Fe_2O_3}$ acts both, as a glass former and as a glass modifier. It is observed that the thermal cum chemical stability of the glasses is increased by higher concentrations of iron. Therefore, the higher iron contents makes the glass network more and more stable, by participating into the glass network formation.

The temperature dependence of DC conductivity confirms the semiconducting nature of the electrical conduction. Typically, the SRPH mechanism is attributed to the electrical conduction. Iron free glass shows a higher conductivity as compared to iron containing $x > 0$ glasses under study. The increase of NBOs resulted in hindrance to the SRPH. Mixed effect of iron and vanadium ions was identified through the variation of AC and DC conductivity. It was confirmed that the glasses behave as non-polar dielectrics. Average σ_{ac} (at 1 kHz and 200 °C) is ≈ 0.07 μS/m, which is comparable to the average DC conductivity at 80 °C ($\sigma_{av} \approx 0.08$ μS/m) for iron containing glasses (i.e., $x > 0$). The average activation energy W_{av} is ≈ 0.70 eV for iron containing glasses (i.e., $x > 0$). The DC conductivity at 180 °C increases consistently with the increase in x and tends to overshoot the conductivity of FVNB0 glass.

Acknowledgments

Authors are thankful to Central Instrumentation Lab (CIL), DCRUST, Murthal for providing FTIR, thermal, and electrical measurement facilities. Arti Yadav is thankful to UGC for providing financial support under Dr. D.S. Kothari Post Doctoral Fellowship Scheme.

References

[1] D.L. Griscom, *Borate Glasses*, New York 1978.

[2] H. Scholze, *Glass*, Springer-Verlag, New York 1991.

[3] N.P. Bansal, R.H. Doremus, *Handbook of Glass Properties*, Academic Press (Indian Reprint) (2006).

[4] M. Regan, C.F. Drake, *Mater. Res. Bull.* **6**, 487 (1971).

[5] M. Nagaswa, H. Watanabe, *UK Patent No. 1358930*, 1971.

[6] P. Jozwiak, J.E. Garbarczyk, M. Wasiucionek, I. Gorzkowska, F. Gendron, A. Mauger, C.M. Julien, *Solid State Ion.* **179**, 46 (2008).

[7] J.E. Garbarczyk, P. Machowski, M. Wasiucionek, W. Jakubowski, *Solid State Ion.* **157**, 269 (2003).

[8] E.I. Kamitsos, M.A. Karakassides, G.D. Chryssikos, *J. Phys. Chem.* **90**, 4528 (1986).

[9] E.I. Kamitsos, G.D. Chryssikos, A.P. Patsis, M.A. Karakassides, *J. Non-Cryst. Solids* **131-133**, 1092 (1991).

[10] L. Bih, M. El Omari, J.M. Réau, M. Haddad, D. Boudlich, A. Yacoubi, A. Nadiri, *Solid State Ion.* **132**, 71 (2000).

[11] A. Langar, N. Sdiri, H. Elhouichet, M. Ferid, *J. Alloys Comp.* **590**, 380 (2014).

[12] A. Mekki, G.D. Khattak, D. Holland, M. Chinkhota, L.E. Wenger, *J. Non-Cryst. Solids* **318**, 193 (2003).

[13] A.A. Bhagat, *J. Non-Cryst. Solids* **226**, 155 (1998).

[14] G. Dalba, P. Fornasini, F. Rocca, F. Monti, *Phys. Chem. Glass.* **41**, 290 (2000).

[15] R.A. Montani, M. Levy, J.L. Souquet, *J. Non-Cryst. Solids* **149**, 249 (1992).

[16] S. Khasa, M.S. Dahiya, A. Agarwal, P. Chand, *J. Mol. Struct.* **1079**, 15 (2015).

[17] S. Khasa, M.S. Dahiya, A. Agarwal, *J. Mol. Struct.* **1086**, 172 (2015).

[18] A. Yadav, S. Khasa, A.M.S. Dahiya, A. Agarwal, P. Chand, *Spectrochim. Acta A* **157**, 129 (2016).

[19] I. Kashif, S.A. Rahman, A.A. Soliman, E.M. Ibrahim, E.K.A. Khalek, A.G. Mostafa, A.M. Sanad, *Physica B* **404**, 3842 (2009).

[20] S. Saetova, A.A. Raskovalov, B.D. Antonov, T.V. Yaroslavtseva, O.G. Reznitskikh, E.V. Zabolotskaya, N.I. Kadyrova, A.A. Telyatnikova, *Ionics* **24**, 1929 (2018).

[21] J.A. Duffy, E.I. Kamitsos, G.D. Chryssikos, A.P. Patsis, *J. Phys. Chem. Solids* **34**, 153 (1993).

[22] V. Dimitrov, T. Komatsu, *J. Ceram. Soc. Jpn.* **107**, 1012 (1999).

[23] J.A. Duffy, M.D. Ingram, *J. Non-Cryst. Solids* **21**, 373 (1976).

[24] J. Duffy, M. Ingram, *J. Am. Chem. Soc.* **93**, 6448 (1971).

[25] R. Iordanova, Y. Dimitriev, V. Dimitrov, D. Klissurski, *J. Non-Cryst. Solids* **167**, 74 (1994).

[26] S. Muthupari, S. Prabhakar, K.J. Rao, *J. Phys. Chem.* **98**, 2646 (1994).

[27] B. Tareev, *Physics of Dielectric Materials*, MIR Publ., Moscow 1979.

[28] N.F. Mott, *J. Non-Cryst. Solids* **1**, 1 (1968).

[29] M.S. Dahiya, S. Khasa, A. Agarwal, *J. Therm. Anal. Calorim.* **123**, 457 (2016).

[30] J.A. Kerr, *Chem. Rev.* **66**, 465 (1966).

[31] A. Yadav, S. Khasa, M.S. Dahiya, S. Dalal, A. Hooda, A. Agarwal, *Phys. Chem. Glass.* **57**, 146 (2016).

[32] B. de B. Darwent, *National Standard Reference Data Series: Bond Dissociation Energies in Simple Molecules*, National Bureau of Standards No. 31, Washington 1970.

[33] I.G. Austin, N.F. Mott, *Adv. Phys.* **18**, 41 (1969).

[34] T.K. Pietrzak, P.P. Michalski, P.E. Kruk, W. Ślubowska, K. Szlachta, P. Duda, J.L. Nowiński, M. Wasiucionek, J.E. Garbarczyk, *Solid State Ion.* **302**, 45 (2017).

[35] L. Murawski, R.J. Barczyński, D. Samatowicz, *Solid State Ion.* **157**, 293 (2003).

[36] V. Naresh, S. Buddhudu, *Ferroelectrics* **437**, 110 (2012).

[37] A. Langar, N. Sdiri, H. Elhouichet, M. Ferid, *J. Alloys Comp.* **590**, 380 (2014).

[38] N. Singh, A. Agarwal, S. Sanghi, *J. Alloys Comp.* **509**, 7543 (2011).

[39] M.P. Kumar, T. Sankarappa, B.V. Kumar, N. Nagaraja, *Solid State Sci.* **11**, 214 (2009).

[40] P.B. Macedo, T. Moynihan, R. Bose, .

[41] A.A. Bhagat, *J. Non-Cryst. Solids* **226**, 155 (1998).

[42] D.D. Ramteke, R.S. Gedam, *Spectrochim. Acta A* **133**, 19 (2014).

[43] A. Hooda, S. Sanghi, A. Agarwal, R. Singh, N. Ahlawat, *J. Appl. Phys.* **112**, 014110 (2012).

[44] S. Sanghi, S. Duhan, A. Agarwal, P. Aghamkar, *Physica B* **405**, 3846 (2010).

IOP Conference Series: Materials Science and Engineering

PAPER • **OPEN ACCESS**

Studies on the effects of braking loads on a Railway Wheel

To cite this article: Aakash D Bansal *et al* 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **691** 012019

View the article online for updates and enhancements.

This content was downloaded from IP address 14.139.251.100 on 19/12/2019 at 04:49

Studies on the effects of braking loads on a Railway Wheel

Aakash D Bansal[1,*], N Ramachandran[2] and V Rastogi[3]

[1,3]Design Centre, Delhi Technological University, Delhi, India
[2]National Institute of Technology Calicut, Kozhikode, Kerala, India
* Corresponding author E-mail: aakash_phd2k17@dtu.ac.in

Abstract. *Indian Railways boasts as the superior modes of transportation in the country and it functions in all rural and urban networks including narrow, meter and broad gauges. Railway accidents due to mechanical failure of wheels are still considered with primary importance by research and development wing. Studies have revealed that the braking of tread zone of wheel results in non-uniform heating and creates severe stresses in railroad car wheels. The existing trend towards high speed trains and larger axle loads has created a need for a better understanding of the technology involved in designing of wheel. All the types of trains including freight and passenger, are generally braked using cast iron brake blocks which adds to the thermal loading on the wheel. Braking application for long time results in high temperatures on the wheel tread and the rim. Thermal loading has sufficient capability to not only initiate cracks but also increases the crack growth rate.*

The focus of the paper is to study the effect of thermal interaction of Micro Alloyed AAR Class-B Grade Steel wheel and blocks during braking and after cool down. FE thermal analysis of a flat shaped wheel profile exposed to stop braking condition is simulated using Abaqus/CAE 6.13. Axle loads, rate of deceleration and brake sizes are taken as parameters to understand the effect of different running conditions on wheel tread surface. The results obtained show the effects of axle load, brake size and rate of deceleration on the peak temperature obtained on the tread surface. The analysis helped in determining the region of the rail wheel which is susceptible to failure due to steep temperature changes

Keywords: *Rail wheel; Tread Braking; Thermal cracking; Finite Element Analysis; Contact Temperature.*

1. Introduction

Indian Railways is one of the world's largest railway networks having 115,000 km of track over an extended route of 65,436 km and comprising 7,172 both rural and urban train stations. But almost 64% of this vast network is not yet electrified. Unlike in electrified systems, the locomotives are braked by conventional brake-block arrangement. Conventionally used Cast Iron brake blocks are prone to thermal deterioration. Almost 80% of the heat generated during braking due to friction between the mating wheel-block interfaces goes into the tread part of wheel. While the wheel surface is gaining heat due to braking, another phenomenon is taking place side by side which allows wheel to lose heat. Natural or forced convection of air allows wheels to lose heat simultaneously, making it susceptible to a cycle of continuous gaining and losing heat and thus leading to thermal fatigue damage. The schematic representation of the phenomenon is shown in Figure 1. Indian wagons have a different wheel web than the ones used in European and Australian continents and are designed to carry heavier axle loads, are braked more frequently and thus make this analysis important from an Indian perspective.

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

Researchers working in related areas have not attempted to explore on thermal analysis of rail wheels. Ekberg et al. [1] studied methods for estimating the residual lifetime of the S-shaped plate rail wheel configuration used across the globe, due to cyclic mechanical and thermal loads. They justified the importance of Rolling Contact Fatigue (RCF) Analysis over Classical Fatigue Analysis and also suggested various failure mechanisms of railway components.

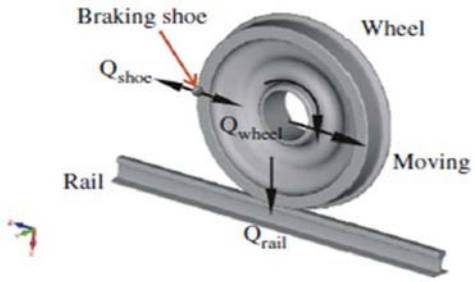


Figure 1. Wheel-Brake and Wheel-Rail interface

Based on the scientific understanding, they developed an engineering model for RCF which defined fatigue in terms of surface and subsurface defects. Peng et al. [2] studied growth of crack in an S-shaped wheel under thermal brake loading and calculated stress intensity factor of cracks using a semi-analytical solution technique that combines an analytical solution and a numerical algorithm to determine fracture strength of the material. The model developed stressed on the heat distributed around the circumference of the tread from friction during braking. Peng et al. [3] combined thermal braking and mechanical contact loads for damage tolerance analysis of a railway wheel and suggested methods to use heat transfer and fracture mechanics tools to predict thermal fatigue-induced cracks. For perfect thermal contact conditions between wheel-rail and wheel-brake interface, a simple semi-analytical numerical method for calculating the wheel tread temperatures was presented by Ertz et al [4] which further suggested an empirical relationship between wheel temperature and average temperature for the constant operating conditions and first contact of the cold wheel respectively could be developed and can lead to the possibilities of thermally induced phase transformations. Teimourimanesh et al. [5] presented a state of the art survey which gave an overview of design of tread braking system with special attention on braking capacity of the railway wheel. Newcomb [6] solved problem of 1-D heat conduction through infinite slab bounded by parallel planes and applied the solution to estimate temperatures during wheel braking at uniform deceleration. Deshpande et al. [7] investigated the problem of wheel tread defect in LHB coaches when the braking system was changed from tread braking to disc braking and concluded that disc brakes are responsible for larger heat generation at the wheel brake interface due to larger torque causing the wheel to spall. Srivastava et al. [8] presented a review of effects of thermal loads on wheel rail contacts and suggested that thermal loads must be considered alongside the evolved defects to completely understand the dynamic material response. Naeimi et al. [9] used FEA to present a coupled thermo-mechanical analysis of wheel-rail contact and evaluate temperature rise and stresses in the contact zone.

2. Thermal Load

When brakes are applied for prolonged duration of time, it results rise in temperature of the tread and rim and is called thermal loading. It has the ability to commence cracks and increase its growth rate. Speed of a train running on a track can be reduced either by Stop Braking or Drag Braking. Stop Braking is the continuous application of brakes till the train is brought to rest. Stop Braking tends to produce large temperatures on the tread surface in a short duration of time. Stop Braking depends on the rate of deceleration. Normally Indian trains decelerate at the rate of 0.5 m/s^2 to 1.2 m/s^2 which may increase up to 1.5 m/s^2 in case of emergency brakes. The paper aims to find the effect of increasing the rate of deceleration on the temperatures attained over the tread surface. Drag Braking or downhill

braking is used to limit the speed of train to a particular value while running down a slope. Potential increase in train speed is detained by applying brakes at constant pressure for longer time. Though drag brakes may be applied for a larger duration, say hours, its overall effect over the wearing of tread surface is less compared to stop braking. Therefore, effect of drag braking is beyond the scope of this paper**.**

Energy required to stop the vehicle is a summation of potential, kinetic and rotational energy. Potential energy is the energy inherited or lost by the vehicle by virtue of its position at rest on hill or at a height. Assuming train to be moving on a leveled track, potential energy will be zero. Kinetic energy is the energy required to bring the vehicle to rest whereas rotational energy is the energy required to reduce the motion of rotating parts to zero and is assumed as 3% of the kinetic energy of the vehicle, based on earlier reports in similar fields. For a wagon with average axle load 10 tonnes per wheel running at 100 km/h if brought to rest at deceleration rate of 0.5 m/s^2, will mean a braking time of 55.5 seconds i.e for almost 56 seconds the tread surface will be exposed to a continuous heat source. In 56 seconds, the train will cover a braking distance of 770 m with the braking energy 3.97 MJ per wheel and initial power of 71.43 kW per wheel.

Now in order to estimate the one-stop temperature rise in the wheel disc interface, an assumption to the location of dissipation of the raking energy has to be made. In the beginning, maximum rise in temperature is observed in the disc; however the disc is rapidly cooled by surrounding components and the natural convection of stream of air. The calculation is done with the assumption that 90% of the total energy generated is retained by the disc. Heat flux into the sides portion of the disc is calculated using the equation in [4]:

$$q=\frac{4P}{\pi(D^2 - d^2)} \tag{1}$$

Temperature rise in a single stop braking is determined using equation:

$$T_{max\emptyset} = \frac{0.527q\sqrt{t}}{\rho ck} + T_{amb} \tag{2}$$

Brake material is assumed to be cast iron with density as 7250 kg/m^3, specific heat capacity 500 J/kg/K and thermal conductivity 58 W/m K.

3. Finite Element Model

In order to conduct the FE thermal analysis of the 920 mm diameter flat shaped of high carbon steel, AAR grade B wheel profile model experiencing stop braking is considered. Only one-half of the wheel is modelled first considering the axial symmetry [10]. The temperature distribution obtained is found to be in complete agreement with the axisymmetric model for the S-shape rail wheel profile by Peng et al [11].

The resultant mesh created using Abaqus®, was with 38060 eight noded linear heat transfer brick element and 44240 nodes. To model the thermal effect of brakes on the interface, a fine mesh in 0.1584 m^2 contact region was is used in the wheel-brake.

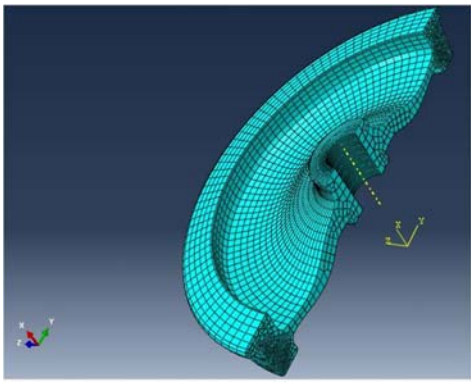


Figure 2. The 3-D mesh of a 920 mm Railway Wheel

4. Material Properties

The material model is assumed to be equivalent to Microalloyed AAR Class B wheel steel and is extracted from Peng et al [11].

Table 1. Material properties of wheel

Yield Strength	360 MPa
Young's Modulus	206 GPa
Poisson's Ratio	0.286
Density	7870 kg/m^3
Specific Heat	490 J/kg °C
Coefficient of Thermal Expansion	14×10^5
Thermal Conductivity	47.5 W/°C m
Free Convection Heat Transfer coefficient	25×10^6 W/ °C m^2

5. Transient Heat Transfer Analysis

Figure 3 represents the temperature attained in the tread region of the railway wheel at different times under stop braking conditions. The train was assumed to be running at a speed of 100 kmph when stop brakes were applied to stop the train in 770 m in duration of 56 seconds. Figure 4. shows the temperature profile obtained for the stop braking case for a duration of 56 seconds. The maximum temperatures of 222 °C are attained in the tread region of the wheel web.

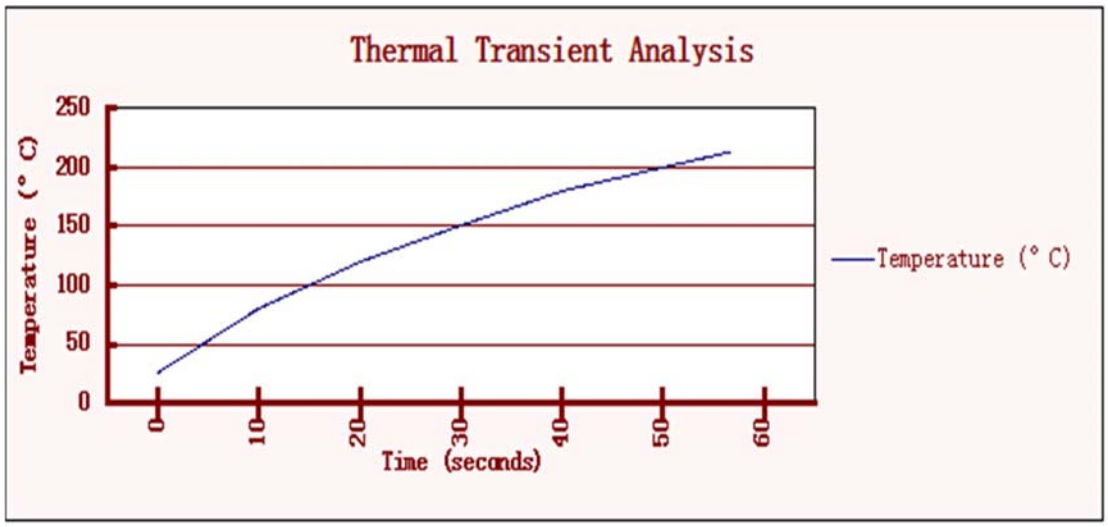


Figure 3. Results of the transient analysis for the maximum temperatures attained in the tread portion of the wheel vs the braking time under stop braking application (average power 71.43 kW per wheel and duration 56 seconds).

6. Results and Discussion

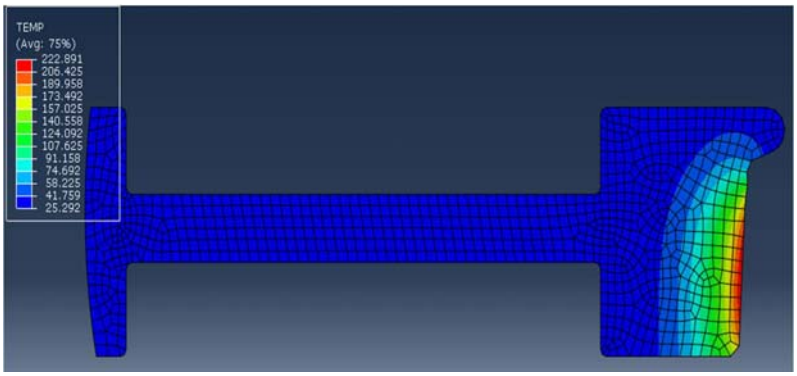


Figure 4. Temperature profile obtained in a flat web shaped railway wheel under stop braking conditions for a duration of 56 s by axisymmetric analysis

.

With an aim to show the effects on the thermal stresses of a Flat-shape plate railway wheel under braking load, various parameters associated with the locomotive in motion were varied and corresponding influence on the temperatures is estimated.

6.1Temperature distribution for stop braking by varying the rate of deceleration

Figure 5 compares the results obtained when rate of deceleration is varied from 0.5 m/s^2 which is the standard rate of deceleration to 1.5 m/s^2 for the case of emergency braking. The maximum temperature obtained in the tread region for a locomotive brought to stop from a speed of 100 kmph was around 220 °C for normal braking whereas temperatures as high as 415 °C were obtained for emergency braking. Approximately 200% more increase in temperature is observed for more severe cases of braking.

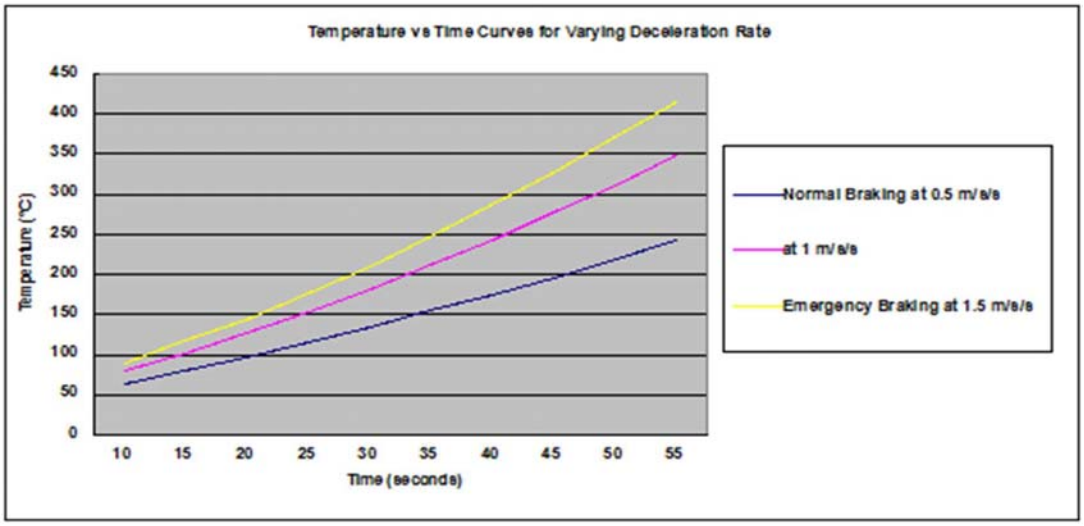


Figure 5. Results of the transient analysis for the maximum temperatures obtained on the wheel tread surface by varying the rate of deceleration from 0.5 m/s^2 standard rate of deceleration to 1.5 m/s^2 for the case of emergency braking.

6.2Temperature distribution for stop braking by varying the axle load capacity

Figure 6 compares the results obtained when tonnage capacity of axle is varied from 15 tonnes which is the standard capacity of a passenger train i.e 7.5 tonnes per wheel to 25 tonnes for a goods train. The maximum temperature obtained for the train having a capacity of 25 tonnes was 390 °C which was approximately 170 % more than the temperatures obtained for the passenger train running under similar conditions. Figure 8 shows similar relationship between heat flux against the time for the same running conditions as above for varying axle loads.

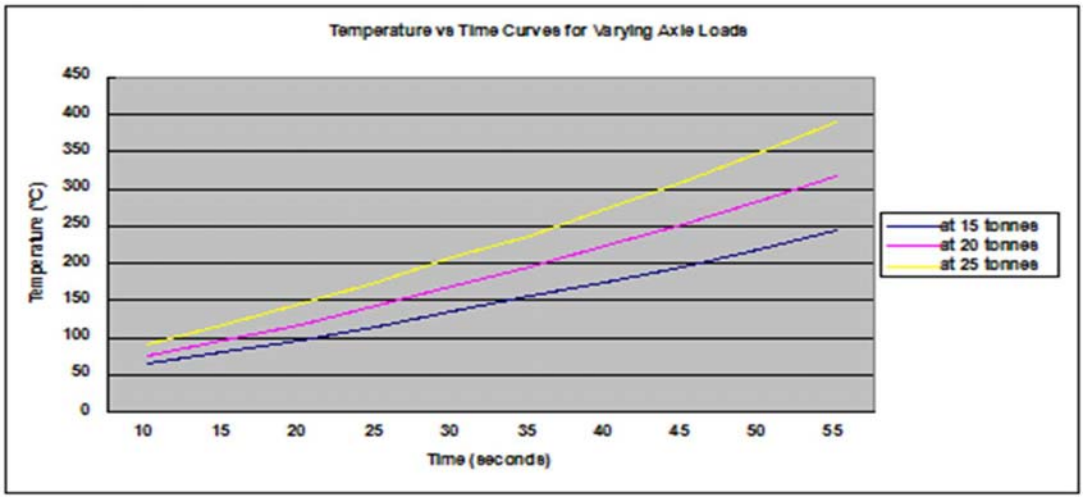


Figure 6. Results of the transient analysis for the maximum temperatures obtained on the wheel tread surface by varying the axle loads from 15 tonnes capacity of the normal passenger locomotive to 25 tonnes capacity of goods and freight wagon.

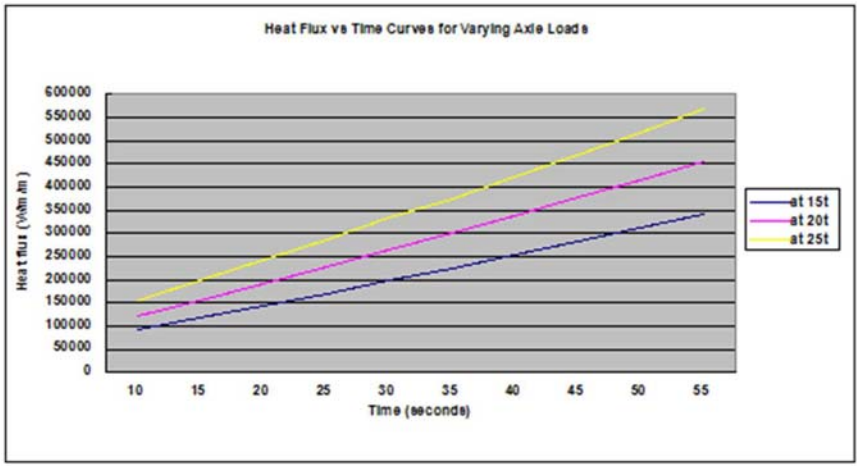


Figure 7. Heat Flux against Time curves for varying axle loads

6.3Temperature distribution for stop braking by varying Wheel-Brake Contact Area

Figure 8 shows the results of changing the Brake contact area by varying the contact angle. Indian railways use brakes of varied contact angles depending upon the maximum speed and load bearing capacity of the locomotive. Braking is accomplished rubbing the brake blocks varying from 0.1584 m^2 to 0.1853 m^2 against the wheel tread surface. Analysis showed 20 to 30 % change in maximum temperature for a change in the contact angle.

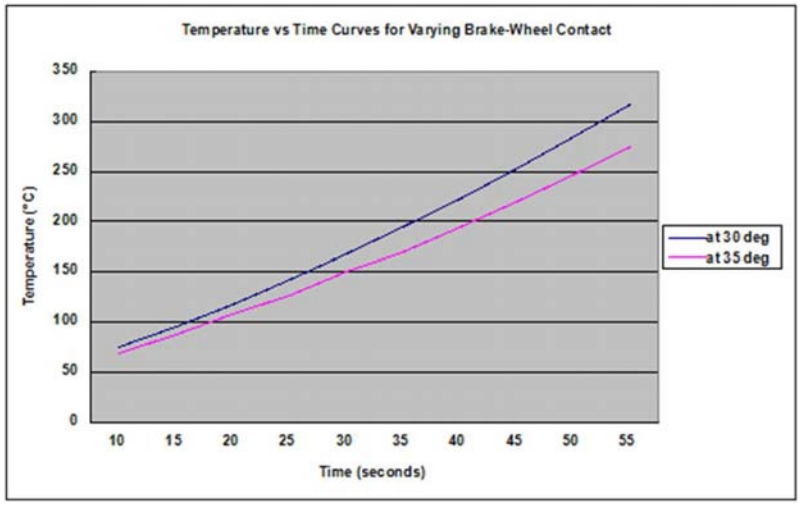


Figure 8. Results of the transient analysis for the maximum temperatures obtained on the wheel tread surface for Brake contact angle of 30° and 35°

7. Conclusion

The studies showed how transient thermal analysis can be used to predict the effect of different parameters of a running train on the maximum temperatures and heat flux obtained on the tread surface. Thus the effect of change in running conditions can be determined. The work indicated that increasing the rate of deceleration has a major impact on the temperatures attained and therefore braking should neither be hard nor frequent. Peak temperature values almost double if the trains are

brought to rest under emergency braking. Considerable changes in temperatures are also observed with a change in Wheel-Brake contact area and load bearing capacity of the locomotive. Issues other than large wear observed in the study like influence of climate, relation between wear and fatigue and rail chill have not been considered. This forms the part for future studies.

8. References

[1] Ekberg A and Kabo E 2005 *Wear* **258(7-8)** pp 1288-1300
[2] Peng D, Jones R and Constable T 2012 *Theor. Appl. Fract. Mech.* **60(1)** pp 10-14
[3] Peng D, Jones R and Constable T 2012 *Eng. Fail. Anal.* **25(1)**, pp 280–290
[4] Ertz M and Knothe K 2002 *Wear* **253(3-4)** pp 498-508
[5] Teimourimanesh S, Lunden R and Vernersson T 2010 *16th International Wheelset Congress (Capetown)*
[6] Newcomb P T 1960 *J. Mech. Sci.* **2(3)** pp 167-177
[7] Deshpande S S, Srikari S, Banthia K V, Jagdeesh K and Chowdhary N 2010 *J. SAS Tech* **9(2)** pp 1–10
[8] Srivastava J P, Sarkar, P K and Ranjan V 2016 *J. Therm. Stress* **39(11)** pp 1389–1418
[9] Naeimi M, Li S, Li Z, Wu J, Petrov R H, Sietsma J and Dollevoet R 2018 *Tribol. Int*. **117** pp 250–260
[10] Orlova A and Boronenko Y *The Anatomy of Railway Vehicle Running Gear* Handbook of Railway Vehicle Dynamics chapter 3
[11] Peng D, Jones R and Constable T 2013 *Eng. Fract. Mech.* **98** pp 1-14

Wireless Personal Communications
https://doi.org/10.1007/s11277-019-07012-7

Theoretical Analysis and Design of High-Stable-Gain Antenna with Ultrawide Band Capabilities and Suppressed Back Radiations

Sandeep Sharma[1,3] · Mainuddin[1] · Binod Kumar Kanaujia[2] · Mukesh Kumar Khandelwal[3]

© Springer Science+Business Media, LLC, part of Springer Nature 2019

Abstract

In this paper, a novel shaped microstrip antenna is proposed for millimeter wave wireless applications. Dual band performance is achieved by embedding the defect in the ground plane. Then, a metallic reflector is used in the near-reactive region beneath the ground plane of the antenna to achieve ultrawideband performance at millimeter wave range. Further, a dielectric superstrate is placed again in the near-reactive region just above the upper face of the antenna which deflects the radiated waves of near-reactive region and influenced the surface wave on the antenna surface and results in the enhanced antenna bandwidth and gain. Simulation and analysis of proposed antenna is done by High Frequency Structure Simulator which is based on Finite Element Method. Again, theoretical analysis of structure is carried out using equivalent circuit theory approach and designed structure is fabricated by standard photolithography process. Simulated, theoretical, and measured results are in accordance with each other.

Keywords Antennas · Microstrip patch antenna · Theoretical analysis · Antenna array

✉ Mukesh Kumar Khandelwal
mukesh.khandelwal89@ieee.org

Sandeep Sharma
sharmadce@yahoo.co.in

Mainuddin
mainuddin@jmi.ac.in

Binod Kumar Kanaujia
bkkanaujia@yahoo.co.in

1 Jamia Millia Islamia, New Delhi 110025, India

2 Department of Computational and Integrative Sciences, Jawaharlal Nehru University, Delhi 110067, India

3 Bhagwan Parshuram Institute of Technology, Rohini-17, New Delhi 110089, India

Published online: 16 December 2019

1 Introduction

No one can remain untouched by Information and Communication Technologies (ICT) in this era where wireless technology appears to reach to its peak. The story of this successful endeavor begins in 1980s with advent of mobile 2G digital technology which increased the voice capacity delivering mobile to the masses by enabling maximum users per channel and thereafter CDMA established the foundation of 3G technology. Mobile 3G introduced high speed internet access for the first time and optimized mobile for data for better and faster connectivity. But a constant compelling need, to deliver faster and better mobile broadband experience, causes an evolution of 4G technology. It provides more data capacity and expanding into new frontiers LTE and LTE advanced [1]. Along with the prominent features of 4G LTE there are few inherent demerits also as it suffers from a restriction on speed and bandwidth for diversified users as well as higher power consumption [2, 3]. Now to gratify the craving of mammoth wireless devices and to cope up with above challenges, a systematic approach is adopted by leaps and bounds to get ahead onto 5G [4, 5]. Researchers are investigating in the frequency band ranging from 27 to 30 GHz and millimeter (MM) wave for 5G [4, 5].

Various antennas have been reported for 5G applications [2, 6–9], however, such type of antenna is very limited in existence which covers both bands of 27–30 GHz and MM wave region. Proposed antenna offers almost stable gain and better radiation characteristics in both above said bands (partial MM wave band). Stable gain is an essential requirement for wide band or ultra wideband (UWB) antennas because radiation characteristics of a practical antenna should be stable within the whole operating band. Various UWB antennas have been reported in last decade [8–12], however, UWB antenna with stable gain for MM wave applications is still a challenging task among researchers. Gain stability of proposed antenna is compared with latest literature and depicted in Table 1. Proposed antenna has least variation within the operating band.

Defected Ground Structure (DGS) has been widely used in microstrip antennas for improving its characteristics [10–13]. UWB characteristics has been achieved by embedding DGS in the ground plane of the antenna and 165.22% of impedance bandwidth has been achieved [12]. However, antenna gain was not stable [12]. DGS also has been successfully employed for dual band operation [14], triple band operation [15], cross-polarization

Table 1 Comparison of gain stability of proposed antenna with literature

References	Bandwidth (GHz)	Measured gain (dBi)			Simulated gain (dBi)		
		Min	Max	Variation	Min	Max	Variation
[6]	**0.8** (3.4–3.8)	3.6	4.1	**0.5**	3.9	4.3	**0.4**
[7]	**1** (5–6)	3	4.1	**1.1**	3.1	5.2	**2.1**
[2]	**4.6** (25.5–30.1)	12	16	**4.0**	12	16.2	**4.2**
[8]	**9** (28–37)	15	17.5	**2.5**	15	17	**2**
[9]	**20** (65–85)	10	15	**5**	8	15	**7**
[10]	**12.8** (1–13.8)	–	–	–	2	8.1	**6.1**
[12]	**19** (2–21)	1	17	**16**	1.5	18	**16.5**
[13]	**7.75** (9.8–17.55)	5	10	**5**	5	12	**7**
Ant_4	**16** (24.5–40.5)	9.1	9.38	**0.28**	9.50	9.53	**0.03**

suppression [13, 14], mutual coupling suppression [16], polarization diversity [16] and many more other applications. DGS also has been analyzed theoretically through equivalent circuit theory approach [13]. Equivalent circuit of the defects of the ground plane can be connected with the equivalent circuit of the main patch through an ideal transformer [13].

In this paper, an antenna is proposed and designed for UWB operation with stable gain characteristics that can act as a promising candidate for potential MM wave applications. Proposed antenna consists a novel shaped dipole type structure embedded with DGS which is fed by an SMA connector through a microstrip line. Further, a metallic reflector is used which reflects back the back-radiations radiated by DGS. Then, a dielectric superstrate is placed above the antenna structure that give rise to the broad impedance bandwidth and a firm stable gain is obtained in desired operating range. Proposed structures are analyzed by Finite Element Method (FEM) using Ansoft™ HFSS v.14. An equivalent circuit model is also presented and analyzed of proposed antennaand theoretical and simulated results are verified with measured results of fabricated prototypes.

2 Antenna Structure and Design

The schematic diagrams of the proposed antennas are shown in Fig. 1. *Ant_1* consist of dual linear tapered structure bounded with dimensions l_1, l_2, l_3, l_4 and l_5 as shown in Fig. 1a. This structure is fed by a 50 Ω microstrip line of length l_{ml} and width w_{ml} which is further connected to an SMA connector. Width w_{ml} of the microstrip line is calculated as

$$w_{ml} = \frac{2h}{\pi}\left[t - 1 - \ln(2t-1) + \frac{(\varepsilon_r - 1)}{2\varepsilon_r}\left\{\ln(t-1) + 0.39 - \frac{0.61}{\varepsilon_r}\right\}\right] \tag{1}$$

where t is defined as

$$t = \frac{377\pi}{2.Z_0\sqrt{\varepsilon_r}} \tag{2}$$

where Z_0 is the characteristic impedance of the microstrip line and taken as 50 Ω.

Length l_{ml} of the microstrip line is taken as the multiple of quarter wavelength at design frequency *freq* which is considered as 28 GHz for this study. Length l_{ml} is calculated as

$$l_{ml} = n \cdot \frac{\lambda_r}{4} = n \cdot \frac{c}{4 \cdot freq \cdot \sqrt{\varepsilon_r}}, \quad where \quad n = 3 \tag{3}$$

Dimension l_2, l_3, l_4, and l_7 are taken as the quarter wavelength at design frequency.

$$l_2 = l_3 = l_4 = l_7 = \lambda_r/4 = \frac{c}{(4 \cdot freq \cdot \sqrt{\varepsilon_r})} \tag{4}$$

Dimension l_1 is taken as the half wavelength at design frequency.

$$l_1 = \lambda_r/2 = \frac{c}{2 \cdot freq \cdot \sqrt{\varepsilon_r}} \tag{5}$$

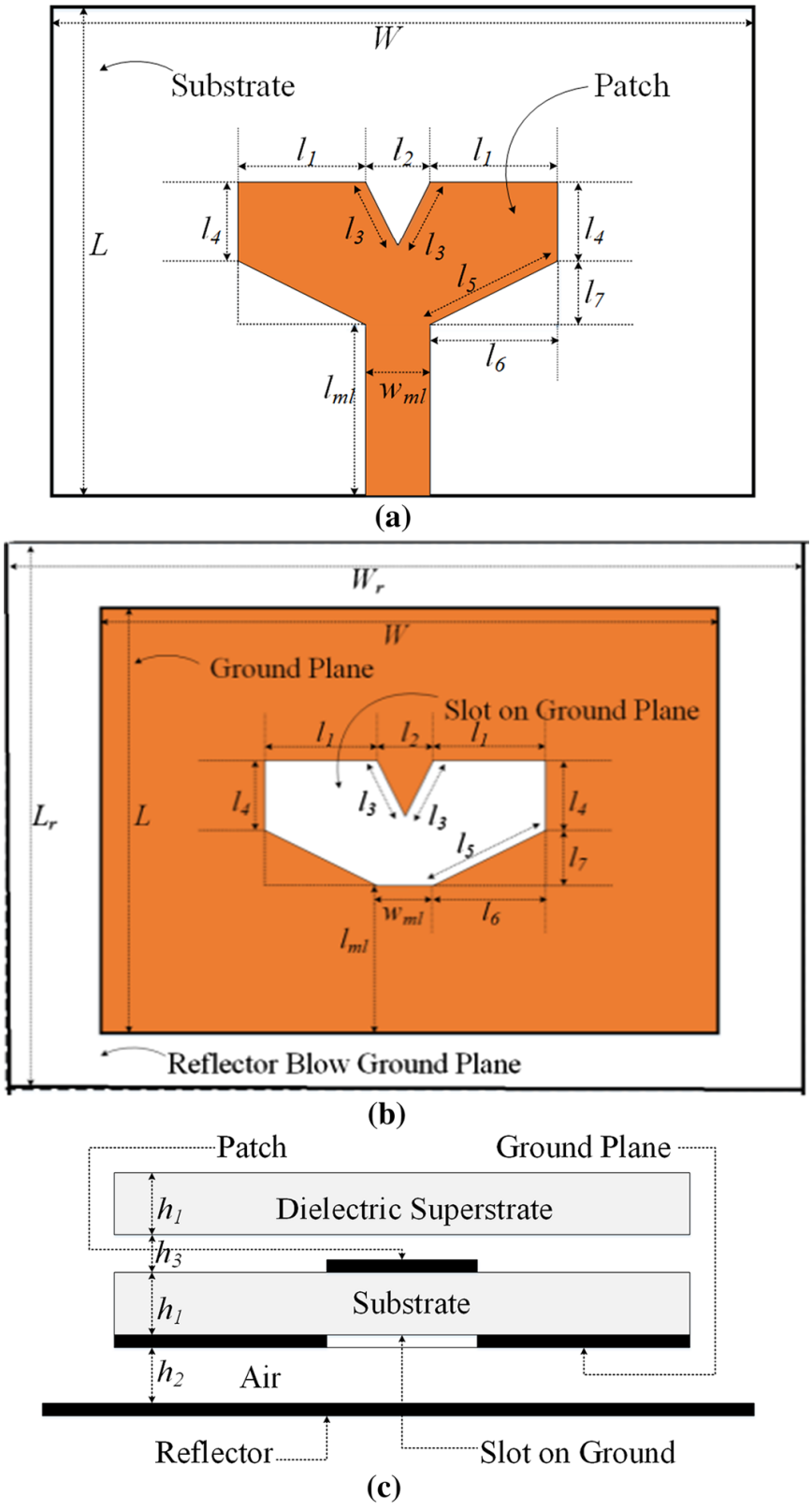


Fig. 1 Schematic diagrams of proposed antennas; **a** top view, **b** bottom view, and **c** side view

Further, a defect of same dimensions of l_1, l_2, l_3, l_4 and l_5 is embedded in the ground plane underneath the patch as shown in Fig. 1b and antenna is referred as *Ant_2*.

In order to improve the gain as well as bandwidth characteristics, this antenna is supported by a rectangular metallic reflector of length L_r and width W_r beneath the ground

planein near reactive region of antenna with an airgap of h_2 distance and is refereed as *Ant_3* as shown in Fig. 1b.

By placing a superstrate of same dimensions as that of substrate with an airgap of h_3 distance improves the gain characteristic and also enhance the maximum peak realized gain, referred as *Ant_4*as shown in Fig. 1c. Dimensions of all parameters are listed in Table 2.

Proposed antenna structure is very simple and easy to design and fabricate. As proposed antenna is fed through microstrip line, it may be embedded with other MMIC devices. Stacking of dielectric superstrate helps to improve the radiation characteristics of the antenna including antenna gain whereas metallic reflector reflects the back radiations in forward direction which further improve the radiation characteristics of the proposed antenna. Presented antenna is fabricated by standard photolithography process and a minimum dimension of 0.1 mm can be fabricated fairly by this process. Thus, frequency corresponding to minimum dimension of 0.1 mm would be considered as maximum frequency for the present study.

3 Theoretical Analysis

Proposed antenna is also analyzed by equivalent circuit theory approach. Equivalent circuit models of proposed structures *Ant_1*, *Ant_2*, *Ant_3*, and *Ant_4*are depicted in Fig. 2b–e respectively. Equivalent circuit model of proposed shaped patch is depicted in Fig. 2a. Proposed patch is fed with a microstrip line which is represented by a tank circuit. Impedance Z_{ml} offered by microstrip line is depending upon series resistance R_{ml}, series inductance L_{ml}, shunt conductance G_{ml} and shunt capacitance C_{ml} and can be determined as [13]:

$$Z_{ml} = \sqrt{(R_{ml} + j \cdot \omega \cdot L_{ml})/\left(G_{ml} + j \cdot \omega \cdot C_{ml}\right)} \tag{6}$$

Impedance of patch Z_{patch} can be calculated as:

$$\left(Z_{Patch}\right)^{-1} = \left(\left(R_p + j\omega L_p + 1/j\omega C_p\right) + 2 \times \left(R_{ST} + 1/j\omega C_{ST}\right)\right)^{-1} + \left(R_{Slot_T} + 1/j\omega C_{Slot_T}\right)^{-1} \tag{7}$$

where R_p, L_p and C_p are the resistance, inductance and capacitance of the conducting part of the patch, respectively. If shaped patch is considered as a rectangle embedded with three triangular slots of dimensions $l_5 \times l_6 \times l_7$, $l_2 \times l_3 \times l_3$, and $l_5 \times l_6 \times l_7$. R_{ST} and C_{ST} are the resistance and capacitance offered by triangular slot of dimensions $l_5 \times l_6 \times l_7$ whereas R_{Slot_T} and C_{Slot_T} are the resistance and capacitance offered by triangular slot of dimensions $l_2 \times l_3 \times l_3$.

Table 2 Dimensions of proposed antenna

Parameter	Dimension (mm)	Parameter	Dimension (mm)
L	15	l_2, l_3 and l_4	1.8
W	18	l_5	3.8
L_r	18	l_6	3.4
W_r	21	l_7	1.8
l_{ml}	5.4	h_1	0.76
w_{ml}	2.4	h_2	1
l_1	3.6	h_3	1

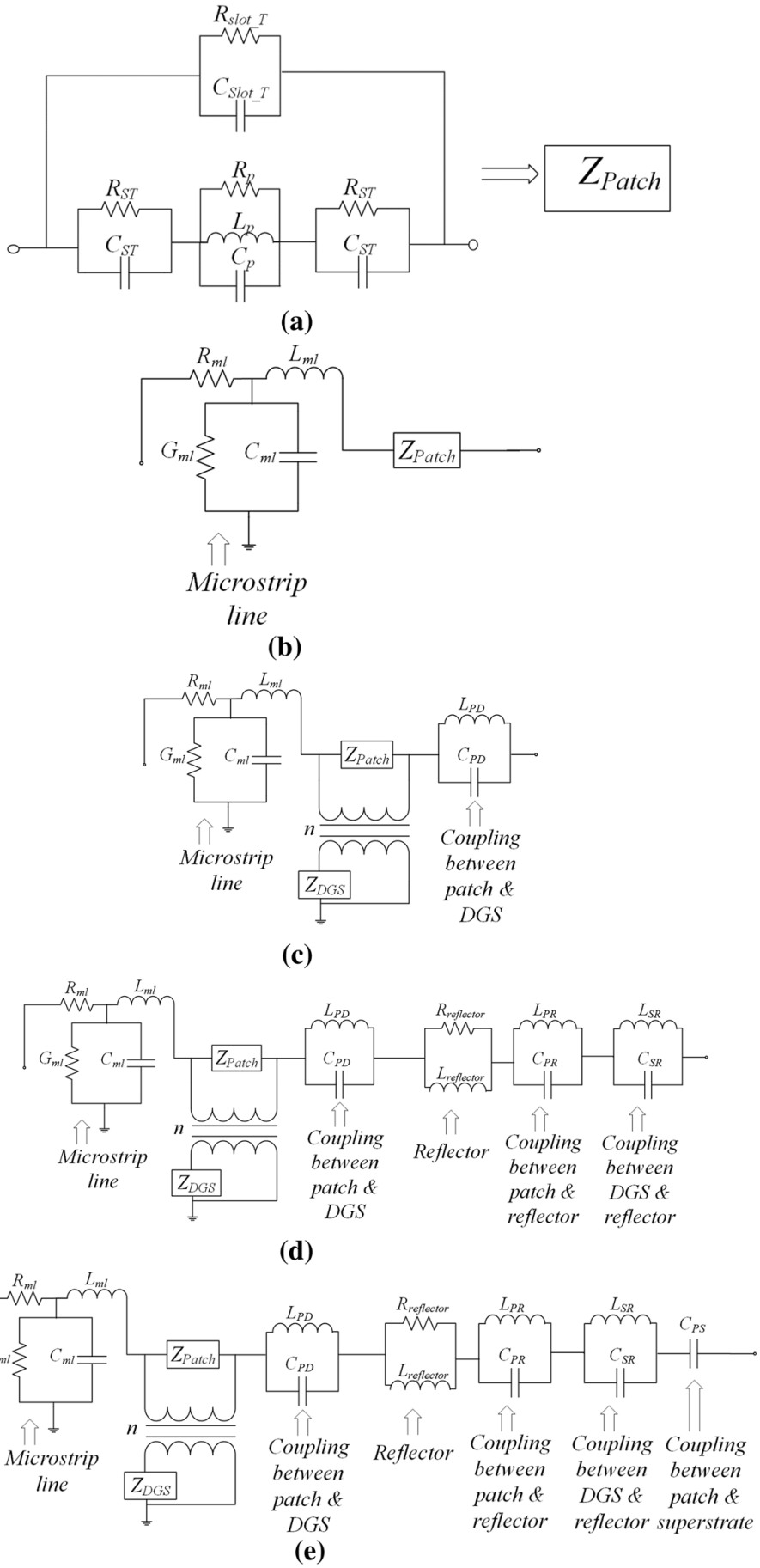


Fig. 2 Equivalent circuit model of, **a** patch only, **b** *Ant_1*, **c** *Ant_2*, **d** *Ant_3*, and **e** *Ant_4*

These parameters can be calculated from [14]. Area of triangular slot can be equated with area of rectangular slot and corresponding rectangular dimensions can be calculated. Further, impedance Z_{slot} offered by a rectangular slot can be determined by the impedance offered by dipole cylinder according to the Babinate's principle [13]:

$$Z_{slot} = \frac{\eta^2}{4 \cdot Z_{cy}} \tag{8}$$

where η and Z_{cy} are the intrinsic impedance of any medium and input impedance of any cylindrical dipole which can be calculated from [13].

Thus, as given in Fig. 2b, input impedance Z_{Ant_1} of *Ant_1* can be calculated as:

$$Z_{Ant_1} = Z_{ml} + Z_{Patch} \tag{9}$$

DGS will provide the complementary effects of patch as both structures have same dimensions and impedance Z_{DGS} offered by DGS can be calculated by the complimentary impedance of Z_{Patch}. Equivalent circuit of DGS is connected with other circuit model with the help of an ideal transformer of turn ratio n. Ideal transformer is lossless so it can be used in circuit model while connecting a DGS circuit to the model. Turn ratio n of transformer is calculated as:

$$n = \sqrt{Z'/Z_{DGS}} \tag{10}$$

where Z' is the equivalent impedance of rest circuit. After resolving the transformer into circuit Z_{DGS} can be replaced by a parallel impedance Z'_{DGS} as:

$$Z'_{DGS} = n \cdot Z_{DGS} \tag{11}$$

If two conducting parts are placed together, coupling capacitance C_{p12} and coupling inductanceL_{p12} will be developed there [16]. Coupling capacitance C_{p12} working between two networks can be determined as [16]:

$$C_{p12} = (-(C_1 + C_2) + \sqrt{(C_1 + C_2)^2 - 4C_1C_2(1 - C_k^{-2})})/2 \tag{12}$$

where C_1, C_2, and C_k are the capacitance of network#1, network#2 and coupling coefficient between both networks respectively. Coupling coefficient C_k is calculated as:

$$C_k = 1/\sqrt{Q_1Q_2} \tag{13}$$

where Q_1 and Q_2 are the quality factors of both networks.

The coupling inductance L_{p12} working between two networks is calculated as:

$$L_{p12} = \left(C_k^2(L_1 + L_2) + \sqrt{C_k^2(L_1 + L_2)^2 + 4C_k^2(1 - C_k^2)L_1L_2}\right)/2(1 - C_k^2) \tag{14}$$

where L_1 and L_2 are the inductance of network#1 and network#2 respectively. Coupling Capacitance C_{PD}, C_{PR}, C_{SR}, and coupling inductance L_{PD}, L_{PR}, L_{SR} are working between patch and DGS, patch and reflector, and DGS and reflector, respectively.C_{PS} is working as a coupling capacitance between superstrate and remaining antenna part because the overlapping part of this dielectric superstrate starts to resonate as its lower face is excited by near-reactive radiated field of the antenna thus a coupling is introduced there between lower excited lower face of the superstrate and patch of the antenna.

Now, input impedance Z_{Ant_2} of *Ant_2* is determined as

$$Z_{Ant_2} = Z_{ml} + \left(1/Z_{Patch} + 1/Z'_{DGS}\right)^{-1} + \left(1/j\omega L_{PD} + j\omega C_{PD}\right)^{-1} \tag{15}$$

Input impedance Z_{Ant_3} is calculated as

$$Z_{Ant_3} = Z_{Ant_2} + \left(1/R_{reflector} + 1/j\omega L_{reflector}\right)^{-1} + \left(1/j\omega L_{PR} + 1/j\omega C_{PR}\right)^{-1} + \left(1/j\omega L_{SR} + 1/j\omega C_{SR}\right)^{-1} \tag{16}$$

Finally, input impedance Z_{Ant_4} is determined as:

$$Z_{Ant_4} = Z_{Ant_3} + \left(1/j\omega C_{PS}\right) \tag{17}$$

Reflection coefficient $|\Gamma|$ is calculated as

$$|\Gamma| = (Z_{Ant} - Z_0)/(Z_{Ant} + Z_0) \tag{18}$$

where Z_0 is taken as 50 Ω and Z_{Ant} can be replaced by Z_{Ant_1}, Z_{Ant_2}, Z_{Ant_3}, and Z_{Ant_4}, respectively. Corresponding return loss S_{11} is obtained as

$$S_{11} = 20\log(|\Gamma|) \tag{19}$$

These formulas are implemented through MATLAB and compared with simulated and measured results in next section.

4 Result and Discussion

Standard photolithography process is used to fabricate theproposed antennasand the fabricated prototypes are shown in Fig. 3. RT-Duriod-5880 substrate of dimensions length L, width W and thickness h with of relative permittivity $\varepsilon_r = 2.2$ is used for the study and fabrication of the proposed antennas. Measurement of proposed antennas are done by Agilent™ Network Analyzer PNA-L series N5230A and Agilent™ Spectrum Analyzer. The radiation characteristics of proposed antennas are measured in Anechoic Chamber.

Figure 4 shows the simulated, measured and theoretical S_{11} variation with frequency of *Ant_1*. *Ant_1* shows 10.63% impedance bandwidth ranging from 26.7 to 29.7 GHz. Simulated and measured gain variations with frequency of *Ant_1* is depicted in Fig. 5. Gain of *Ant_1* is varying from 6.5 to 7.6 dBi within the operating band with peak value of 7.6 dBi at 27.7 GHz.

Further, DGS is embedded in the ground plane and the effect of DGS can be observed clearly from the variation of reflection coefficient (S_{11}) with frequency of *Ant _2* which is shown in Fig. 6. With the effect of DGS *Ant_2* acts like a dual band antenna with an impedance bandwidth of 2.8 GHz ranging from 27.7 to 31 GHz in the 5G proposed band (27–30 GHz) and another band of 3.1 GHz from 34.7 to 37.8 GHz along with a wide notch band of 3.7 GHz. Current is distributed over both the parts of the antenna; patch as well as ground. When a defect is introduced in the ground plane, current lines are disturbed and starts to flow along the periphery of the defect thus it starts to give the effects same as in the case of slot antenna and DGS starts to give an equivalent impedance according to the complimentary theory of Babinate's principle. Coupling is also introduced there between DGS and main patch of the antenna. Thus, a coupling capacitance and coupling inductance

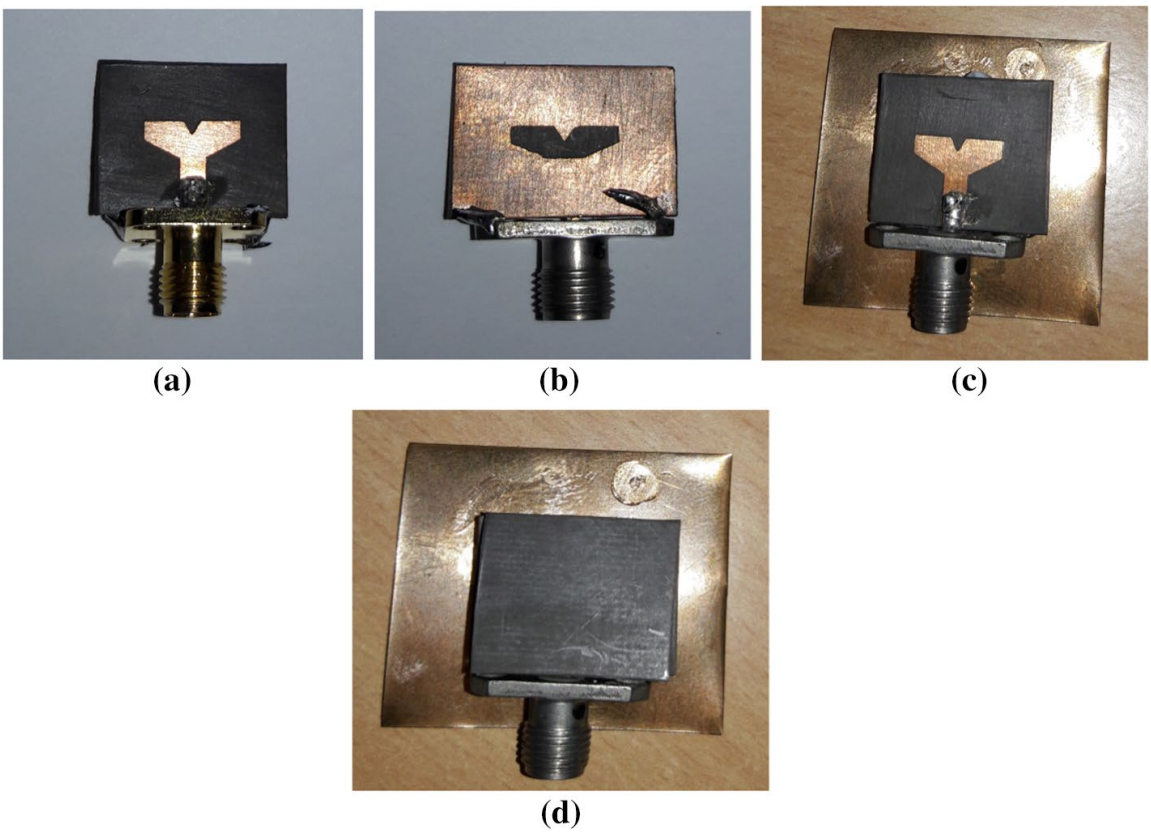


Fig. 3 Prototypes of fabricated antennas; **a** top view of *Ant_1*, **b** ground plane of *Ant_2*, **c** *Ant _3* with reflector and **d** *Ant_4* with reflector and superstrate

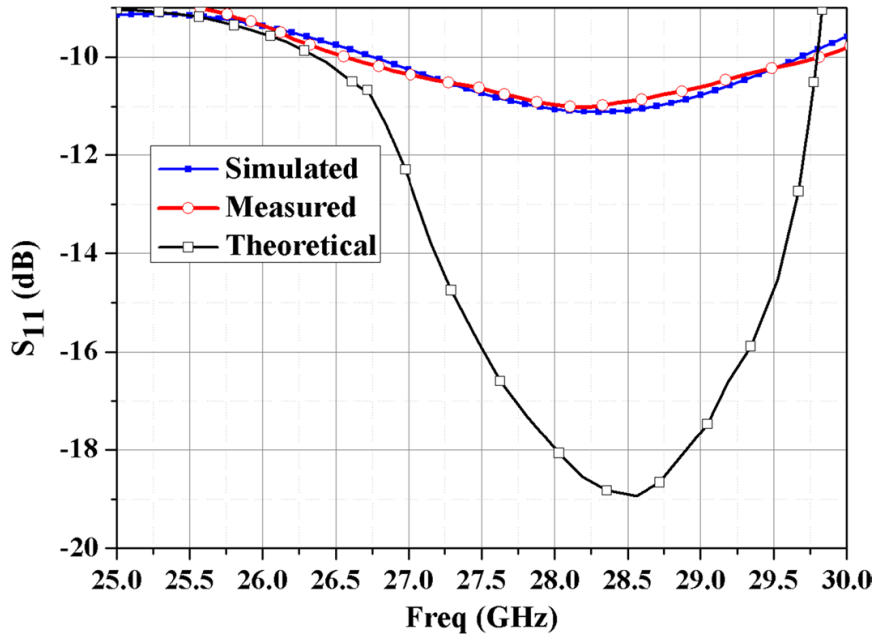


Fig. 4 Theoretical, simulated and measured S_{11} variation with frequency of *Ant_1*

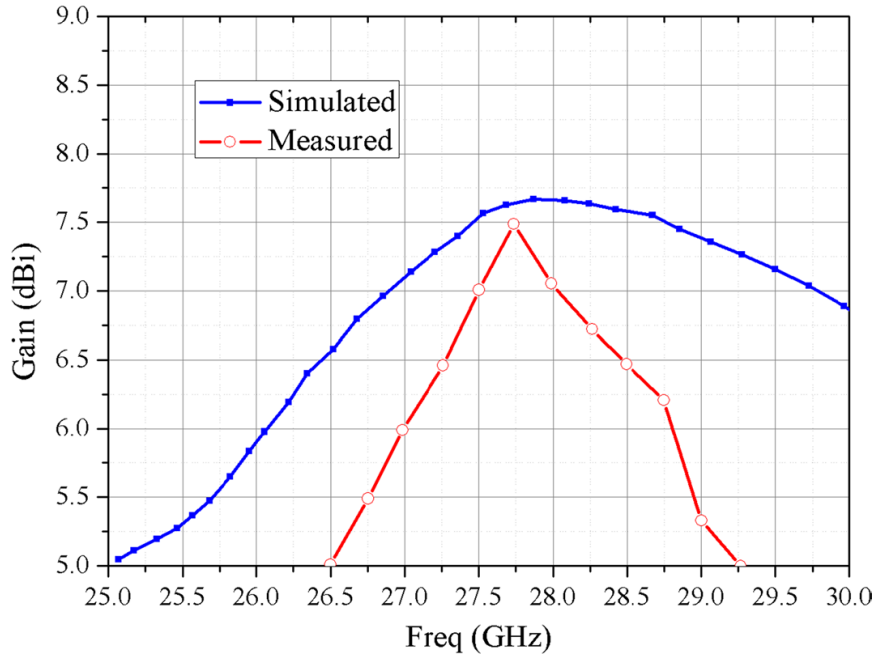


Fig. 5 Simulated and measured gain variations with frequency of the *Ant_1*

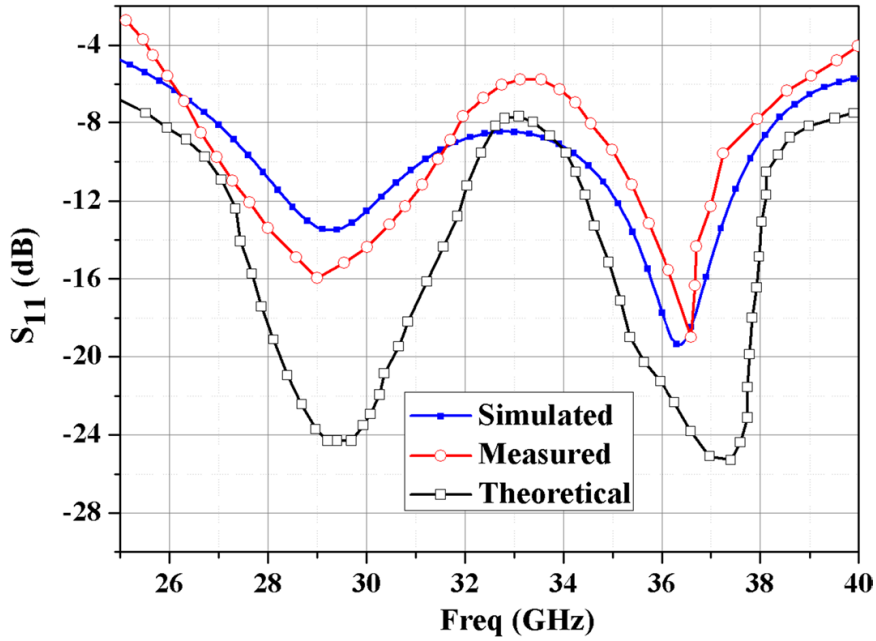


Fig. 6 Theoretical, simulated and measured S_{11} variation with frequency of *Ant_2*

is also introduced there. Thus, reflection coefficient of the structure is changed due to this equivalent impedance of DGS and its coupling with patch.

It is observed that lower and higher cut-off frequencies are in good agreement for simulated and theoretical results. Because theoretical analysis is based on equivalent circuit theory approach which gives only an approximation in case of antennas on the other hand simulated results are based on Finite Element Method (FEM) approach. As both approaches for simulation and theoretical analysis are different and have different loss-considerations thus a difference between simulated and theoretical results are observed. This difference is

occurred just because of different loss-considerations between both techniques. However, lower and higher cutoff frequencies are approximately same for both simulated and theoretical approaches.

Figure 7 shows the gain variation with frequency of *Ant_2* and maximum peak realized gain of 7.19 dBi is achieved at 29.2 GHz in the desired frequency range and 7.6 dBi at 36 GHz in second band. Due to the coupling effect of DGS with patch, radiation properties also changed thus this variation in antenna gain is achieved in comparison of *Ant_1*.

By placing the conducting slab (reflector) of thickness 0.5 mm beneath the *Ant_2* which is referred as *Ant_3*; significantly transforms a dual band of *Ant_2* into an ultrawide-band ranging from 26.6 to 38 GHz. Return loss level of *Ant_3* is also improved and −20 dB impedance bandwidth is achieved from 28 to 29 GHz as shown in Fig. 8. −10 dB bandwidth indicates that about 90% power is going in forward direction, whereas, -20 dB impedance bandwidth indicates that about 99% power is going in forward direction within the band. This conducting reflector is placed in the near-reactive region of the antenna thus it influences the radiated waves as well as the surface waves of the antenna. Thus this conducting reflector not only minimizes a portion of reflected waves on the surface of the antenna and helps to achieve a wide bandwidth but also reflect back the backward radiated waves in forward direction which causes gain enhancement such that an almost stable peak realized gain of 8.8 dBi is obtained in desired frequency band (27–30 GHz) as well as a maximum peak realized gain of 9.3 dBi at 35.5 GHz as the Fig. 9 represents.

Figure 10 shows the 3-d radiation pattern of radiated waves of proposed structures at 28 GHz. Only single working frequency is considered in this figure for comparison purpose. *Ant_1* has its almost radiations in forward direction due to the use of regular ground plane (without defect) as shown in Fig. 10a but *Ant_2* shows backward radiations due to incorporation of DGS as shown in Fig. 10b. Backward radiations of *Ant_2* are reflected back in the forward direction by using a conducting reflector and *Ant_3* shows two main lobes in +Z direction as shown in Fig. 10c which also results in enhanced gain. Dual beam antenna could be used in scanning radar applications. Radiation pattern of *Ant_4* is shown

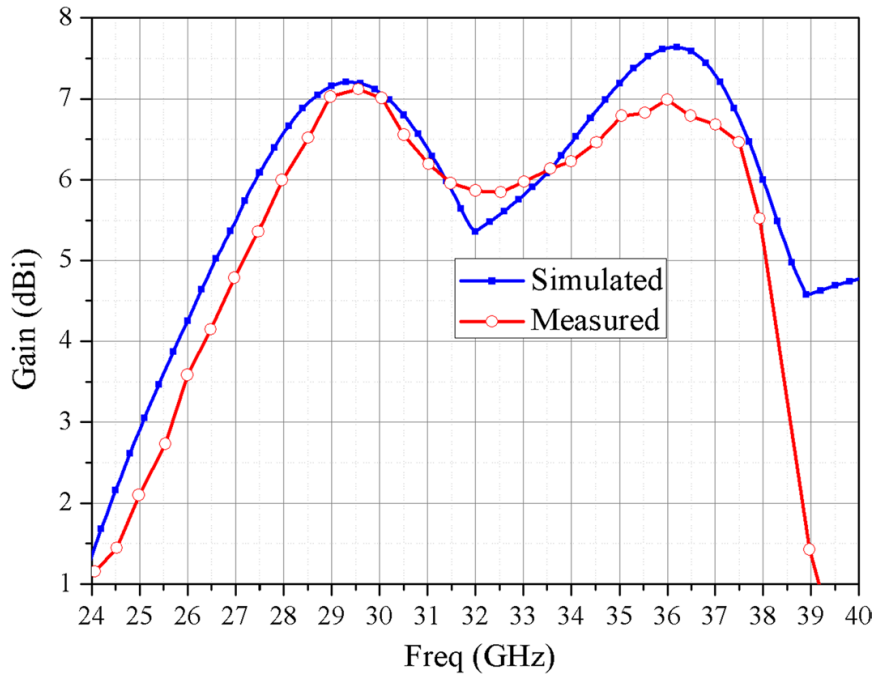


Fig. 7 Simulated and measured gain variations with frequency of the *Ant_2*

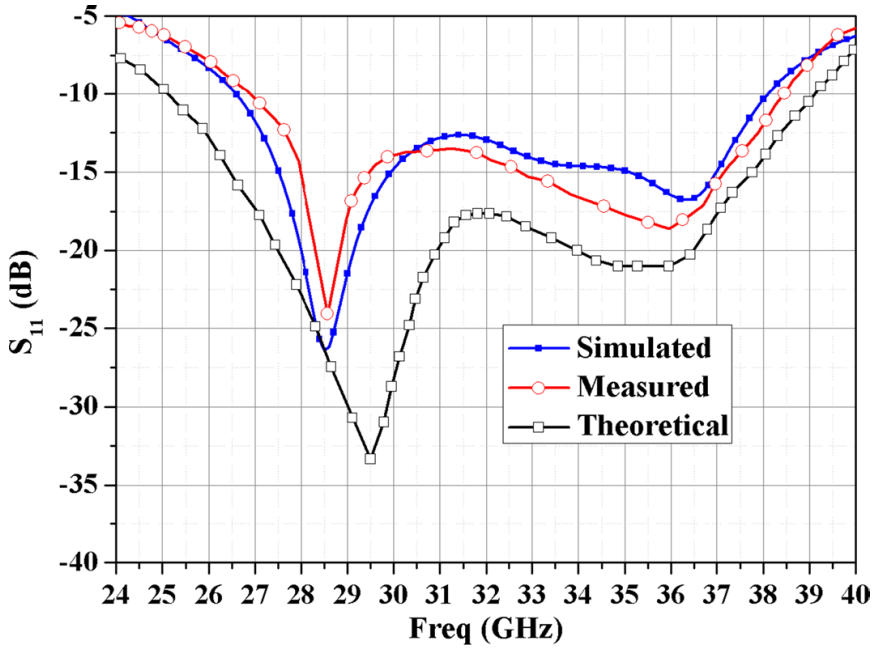


Fig. 8 Theoretical, simulated and measured S_{11} variation with frequency of *Ant_3*

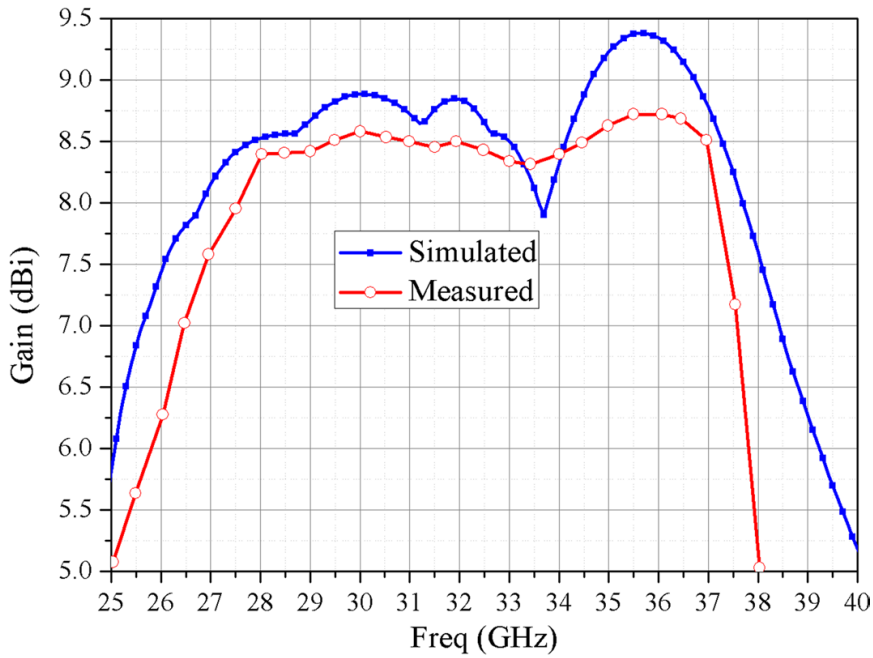


Fig. 9 Simulated and measured gain variations with frequency of the *Ant_3*

in Fig. 10d. Use of dielectric superstrate above the patch of the antenna makes antenna radiations more stable and it can be observed by comparing the Fig. 10c and d. In Fig. 10d, variation between the two major lobes is minimized while comparing with *Ant_3*. For the comparison purpose, radiation characteristics are shown on a common operating frequency i.e. 28 GHz which is also considered as design frequency for the proposed structure. Simulated results are obtained from HFSS simulator whereas theoretical results are obtained from the proposed equations through equivalent circuit model using MATLAB. Measured results are experimental results obtained from the experimental measurement of

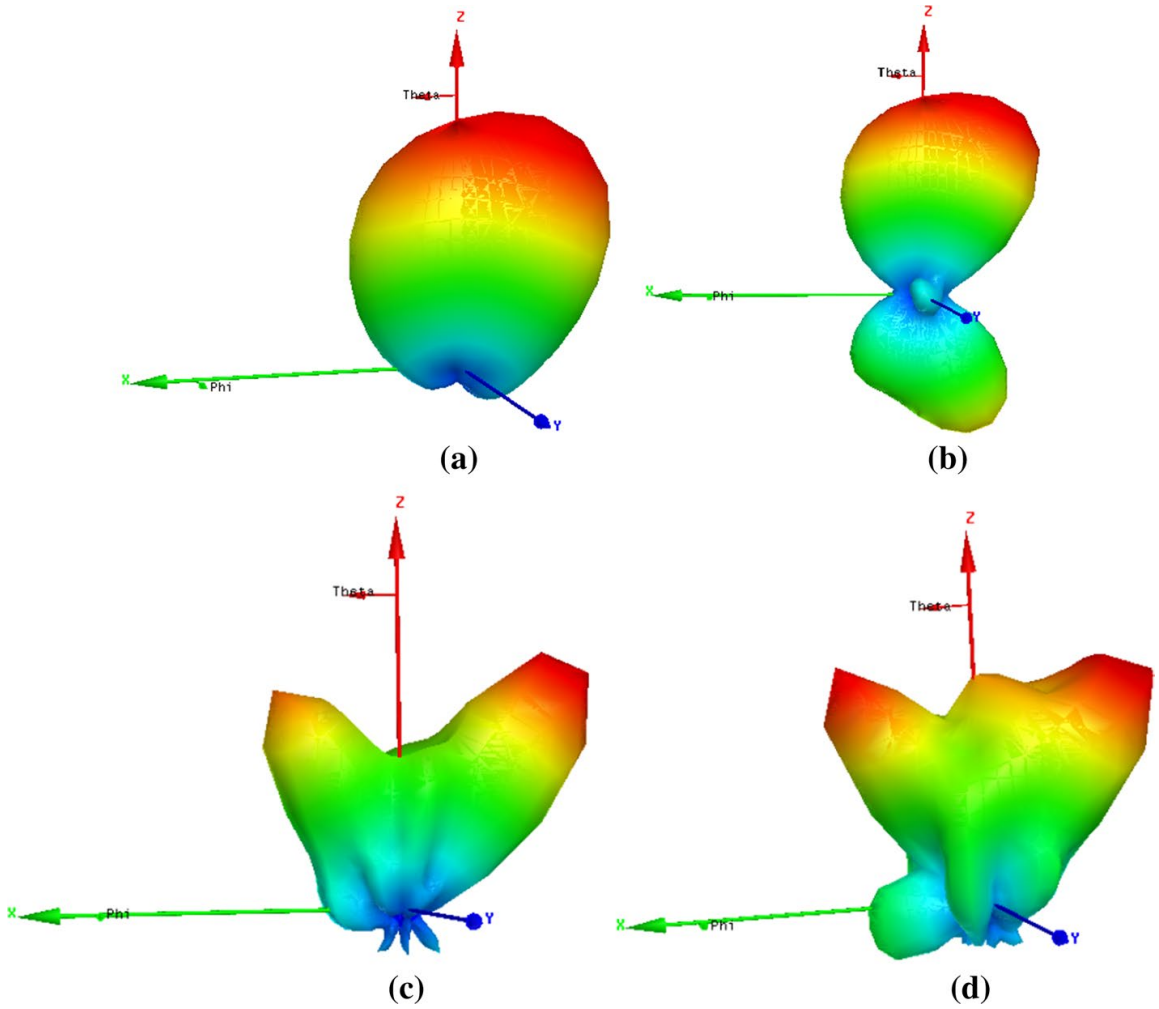


Fig. 10 3-d Radiation patterns at 28 GHz of **a** *Ant_1*, **b** *Ant_2*, **c** *Ant_3*, and **d** *Ant_4*

proposed fabricated prototypes. Although, all three results (i.e. simulated, theoretical and measured) are in good match however a small difference is observed among all the three results. Because simulated results are obtained from HFSS which is based on Finite Element Method (FEM) whereas theoretical results are obtained from equivalent circuit theory approach which give only an equivalent estimation of working model to circuit model. On the other hand, measured results include few losses i.e. fabrication loss, soldering loss, and SMA connector loss. Therefore, a small but acceptable difference is observed among simulated, theoretical and measured results.

A dielectric superstrate, identical to antenna-substrate is introduced above the patch of the antenna with an air gap h_3 shows impedance bandwidth from 24.5 to 41 GHz and a significant improvement can be seen in antenna gain using a superstrate. Almost stable gain of 9.52 dBi is achieved within the whole impedance bandwidth. Simulated gain is varying from 9.50 to 9.53 dBi whereas the value of measured gain is varying from 9.1 to 9.38 dBi. Thus, stable gain with minimum variations of the values of 0.03 and 0.28 dBi is achieved in the simulated and measured results of *Ant_4*, respectively. This dielectric superstrate is also placed in the near-reactive region of the antenna which perturbs the immediate radiated field of the antenna and get excited. The lower face of this dielectric superstrate is excited from the near-reactive radiated field of the antenna thus coupling is introduced there between the patch and the lower layer of the dielectric superstrate. This coupling could be balanced by a coupling capacitor in

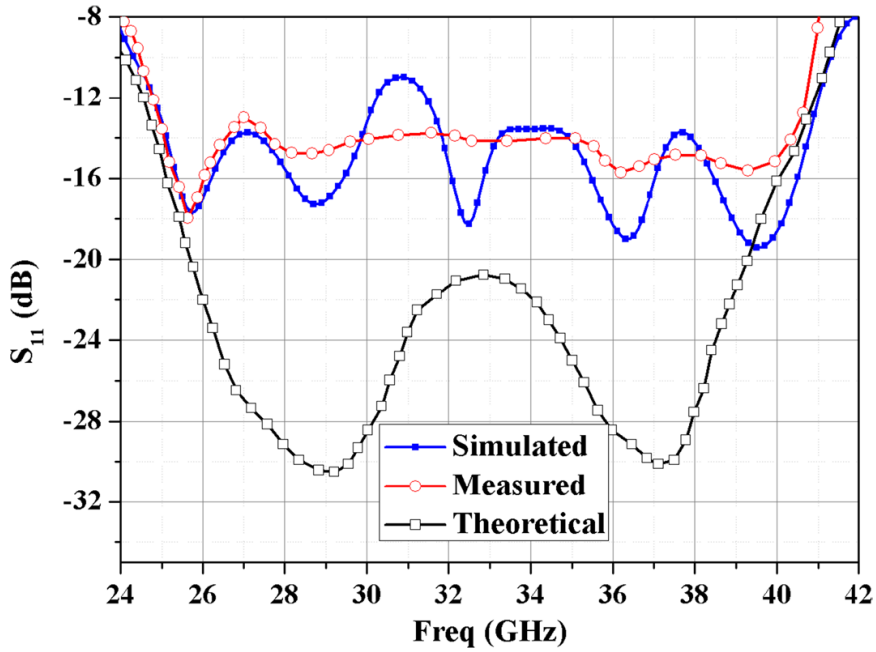


Fig. 11 Theoretical, simulated and measured S_{11} variation with frequency of *Ant_4*

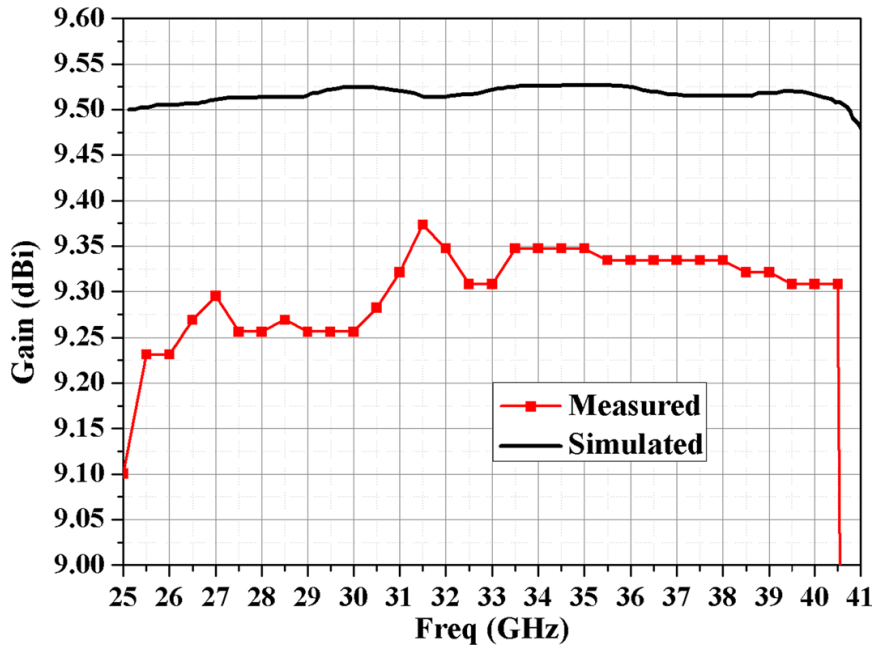


Fig. 12 Simulated and measured gain variations with frequency of the *Ant_4*

equivalent circuit model of the antenna. Now, this excited lower face of the superstrate starts to control the radiation from main antenna part and results in a stable gain. S_{11} and gain variations with frequency of *Ant_4* are shown in Figs. 11 and 12, respectively. All theoretical, simulated and measured results are in good validation to each other and summarized in Table 3. Current distribution of the proposed antenna structures at 28 GHz is depicted in Fig. 13. It is clearly observed that current density causing for radiation is higher for *Ant_1* as it doesn't have any defect in the ground plane. After embedding the defect in the ground plane underneath the main radiating patch, current is coupled to the defect of the ground from the main patch which is observable in Fig. 13b. After placing the metallic reflector underneath the ground plane, it

Table 3 Result summary of presented study

S. nos.	Antenna	Structure specification	Structure	Bandwidth	Gain
1	*Ant_1*	Patch with regular ground		26.7–29.7 GHz	6.5–7.6 dBi
2	*Ant_2*	Patch with DGS		27.7–31 GHz and 34.7–37.8 GHz	7.19 dBi and 7.6 dBi
3	*Ant_3*	Patch with DGS and reflector		26.6–38 GHz	Stable gain of 8.8 dBi within 27–30 GHz band with peak gain of 9.3 GHz at 35.5 GHz
4	*Ant_4*	Patch with DGS, reflector and superstrate		24.5–40.5 GHz	Stable gain of 9.52 dBi

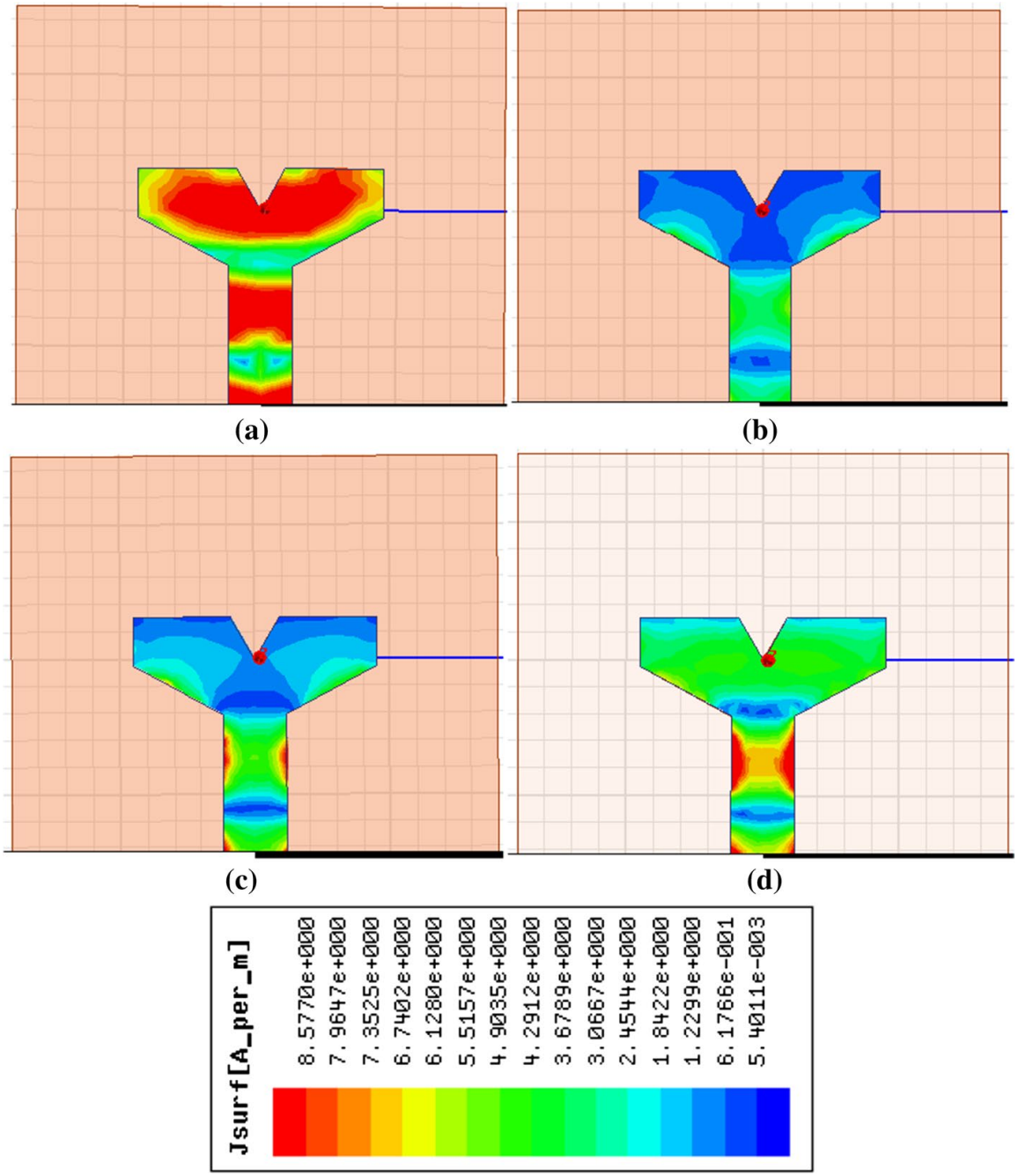


Fig. 13 At 28 GHz current distribution of proposed antennas of **a** *Ant_1*, **b** *Ant_2*, **c** *Ant_3*, and **d** *Ant_4*

starts to reflect back the radiations in forward direction which again increase the current density at the main radiating patch which is further enhanced after using the dielectric superstrate above the radiating patch as shown in Fig. 13c and d respectively.

5 Conclusion

An UWB antenna with stable and high gain for MM wave applications is designed, analyzed and presented. Theoretical analysis is also presented for a shaped microstrip antenna embedded with DGS, metallic reflector and dielectric superstrate. The results of

fabricated antenna show a good operating bandwidth from 24.5 to 41 GHz (16.5 GHz). The simulated and theoretical results are in fine acceptance with measured values. In the 5G proposed band (27–30 GHz), a flat stable high gain of 9.25 dBi is achieved. The proposed antenna is appropriate to use for ultrawide band applications and is best suited for 5G and MM wave wireless applications because of its low return loss and stable high gain within the band.

References

1. Haraz, O. M., Ali, M. M. M., Alshebeili, S., & Sebak, A. R. (2015). Design of a 28/38 GHz dual band printed slot antenna for the future 5G mobile communication networks. *IEEE International Symposium on Antenna and Propagation and USNC/URSI National Radio Science Meeting, 10,* 1532–1533.
2. Haraz, O. M., Elboushi, A., Alshebeili, S., & Sebak, A. R. (2014). Dense dielectric patch array antenna with improved radiation characteristics using EGB ground structure and dielectric superstrate for future 5G cellular networks. *IEEE Access, 2,* 909–913.
3. Tefiku, F. (2008). A dual-band GPS/BT antenna with low coupling to a cellular antenna for mobile phones. In: *IEEE antennas and propagation society international symposium*.
4. Wang, C.-X., et al. (2014). Cellular architecture and key technologies for 5G wireless communication networks. *IEEE Communications Magazine, 52,* 122–130.
5. Tan, H., Li, W., Wang, T., Fang, J., & Feng, Z. (2015). The analysis on the candidate frequency bands of future mobile communication system. *Communications China, 12,* 140–149.
6. Saxena, S., et al. (2018). MIMO antenna with built-in circular shaped isolator for sub-6 GHz 5G applications. *Electronics Letters, 54,* 478–480.
7. Lai, H. W., & Wong, H. (2015). Substrate integrated magneto-electric dipole antenna for 5G Wi-Fi. *IEEE Transactions on Antennas and Propagation, 63,* 870–874.
8. Dadgarpour, A., Sorkherizi, M. S., & Kishk, A. A. (2016). Wideband low-loss magneto-electric dipole antenna for 5G wireless network with gain enhancement using meta lens and gap waveguide technology feeding. *IEEE Transactions on Antennas and Propagation, 64,* 5094–5101.
9. Laurinaho, J. A., et al. (2016). 2-D beam-steerable integrated lens antenna system for 5G *E*-band access and backhaul. *IEEE Transactions on Microwave Theory and Techniques, 64,* 2244–2255.
10. Saha, P., et al. (2015). Design and analysis of UWB circular ring two element microstrip patch antenna array with notched band for modern wireless applications. *Microwave and Optical Technology Letters, 57,* 2067–2072.
11. Khandelwal, M. K., Kanaujia, B. K., & Gautam, A. K. (2013). Low profile UWB log-periodic dipole antenna for wireless communication with notched band. *Microwave and Optical Technology Letters, 55,* 2901–2906.
12. Khandelwal, M. K., Kanaujia, B. K., Dwari, S., & Kumar, S. (2014). Bandwidth enhancement and cross-polarization suppression in ultra-wideband microstrip antenna with defected ground plane. *Microwave and Optical Technology Letters, 56,* 2141–2146.
13. Khandelwal, M. K., Kanaujia, B. K., Dwari, S., & Kumar, S. (2014). Analysis and design of microstrip-line-fed antenna with defected ground structure for Ku band applications. *International Journal of Electronics and Communication, 68,* 951–957.
14. Khandelwal, M. K., Kanaujia, B. K., Dwari, S., & Kumar, S. (2015). Analysis and design of dual band compact stacked microstrip patch antenna with defected ground structure for WLAN/WiMax applications. *International Journal of Electronics and Communication., 69,* 39–47.
15. Khandelwal, M. K., Kanaujia, B. K., Dwari, S., Kumar, S., & Gautam, A. K. (2015). Triple band circularly polarized microstrip antenna with defected ground structure for wireless applications. *International Journal of Microwave and Wireless Technologies, 08,* 943–953.
16. Khandelwal, M. K., Kumar, S., & Kanaujia, B. K. (2018). Design, modeling and analysis of dual feed defected ground microstrip patch antenna with wide axial ratio bandwidth. *Journal of Computational Electronics, 17,* 1019–1028.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Sandeep Sharma received his B.E. in Electronics and Communication Engineering from Delhi College of Engineering in 2003, M.Tech. in Digital Communication in 2011 from AIACTR and pursuing Ph.D. degree from Jamia Millia Islamia, Delhi. Mr. Sharma is working as an Assistant Professor in Bhagwan Parshuram Institute of Technology, Delhi, India. It is affiliated to GGSIP University (Government of Delhi, India). He has published one paper in IEEE Conference and two in other International Conferences. His current research interest focuses on microstrip antennas, defected ground structure, metamaterials and UWB antennas.



Mainuddin received Bachelor's degree in Electronics and Communication Engineering from Jamia Millia Islamia in 1994 with a CPI of 96.49 (First position with Honors.). He received M.E. degree from the Department of Electronics and Communication Engineering, Delhi College of Engineering in 2003, securing First position with Honors (86.16%). He completed his Ph.D. in 2008 from Jamia Millia Islamia, New Delhi. He is presently working as a Professor and Head at the department of Electronics and Communication Engineering, Jamia Millia Islamia, New Delhi. He worked as a Senior Research Scientist at Laser Science and Technology Centre, DRDO, New Delhi for more than 16 years. His research interests include Optoelectronics, Optical diagnostics andSensors. He has published 20 research papers in archival journals with high impact factor like IEEE Transactions, Optical Research, Optics and Laser Technology, Quantum Electronics, Sensors and Actuators: B, Infrared Physics and Technology and Measurement Science and Technology etc. He has authored a book on "Sensors and Measurement Techniques for Chemical Gas Lasers" by International Frequency Sensor Association (IFSA) Publishing, Barcelona, Spain and contributed seven book chapters published by Springer, Cambridge University Press and Nova Publishers. In addition, he has presented 20 papers in International refereed conferences and 35 national conferences/symposiums. He serves as reviewer (on invitation from Editor) for IEEE Transaction on Instrumentation and Measurement, Instrumentation Science and Technology and Optics and Laser Technology.



Binod Kumar Kanaujia is currently working as Professor in the School of Computational and Integrative Science, Jawaharlal Nehru University, Delhi-110067, India since 01-Aug-2016. Previously Dr. Kanaujia was working as a Professor in the Department of Electronics and Communication Engineering in Ambedkar Institute of Advanced Communication Technologies and Research (formerly Ambedkar Institute of Technology), Geeta Colony, Delhi from 17-Aug-2012. Dr. Kanaujia joined this institute as Assistant Professor in 2008 through selection by Union Public Service Commission, New Delhi, India and promoted to Associate Professor on 21-Feb-2008. He served on various key portfolios i.e. Head of Department, In-charge Central Library, Head of Office, etc. Before joining this institute, he had served in the M.J.P. Rohilkhand University, Bareilly, India as Reader in the Department of Electronics & Communication Engineering and also as Head of the Department. He has been an active member of Academic Council and Executive Council of the M.J.P. Rohilkhand University and played a vital role in academic reforms. Prior, to his career in academics, Dr. Kanaujia had worked as Executive Engineer in the R&D division of M/s UPTRON India Ltd. Dr. Kanaujia had completed his B.Tech. in Electronics Engineering from KNIT Sultanpur, India in 1994. He did his

M.Tech. and Ph.D. in 1998 and 2004; respectively from Department of Electronics Engineering, Indian Institute of Technology Banaras Hindu University, Varanasi, India. He has been awarded Junior Research Fellowship by UGC Delhi in the year 2001–2002 for his outstanding work in electronics field. He has keen research interest in design and modelling of microstrip antenna, dielectric resonator antenna, left handed metamaterial microstrip antenna, shorted microstrip antenna, ultra wideband antennas, reconfigurable and circular polarized antenna for wireless communication. He has been credited to publish more than 105 research papers with more than 200 citations with h-index of 10 in peer-reviewed journals and conferences. He had supervised 45 M.Tech. and 03 Ph.D. research scholars in the field of microwave engineering. He is a reviewer of several journals of international repute i.e. IET Microwaves, Antennas and Propagation, IEEE Antennas and Wireless Propagation Letters, Wireless Personal Communications, Journal of Electromagnetic Wave and Application, Indian Journal of Radio and Space Physics, IETE Technical Review, International Journal of Electronics, International Journal of Engineering Science, IEEE Transactions on Antennas and Propagation, AEU-International Journal of Electronics and Communication, International Journal of Microwave and Wireless Technologies, etc. Dr. Kanaujia had successfully executed 04 research projects sponsored by several agencies of Government of India i.e. DRDO, DST, AICTE and ISRO. He is also a member of several academic and professional bodies i.e. IEEE, Institution of Engineers (India), Indian Society for Technical Education and The Institute of Electronics and Telecommunication Engineers of India.



Mukesh Kumar Khandelwal was born in Uttar Pradesh, India and received his B.Tech. degree in Electronics & Communication Engineering in 2010 from UPTU, M.Tech degree in Digital Communication from GGSIP University, Delhi, India in 2012 and Ph.D Degree in Electronics Engineering from Indian Institute of Technology ISM, Dhanbad, India in 2015. Dr Khandelwal is working as an Associate Professor in the Department of Electronics & Communication Engineering, Bhagwan Parshuram Institute of Technology, Delhi India which is affiliated to Guru Gobind Singh Indraprastha University Delhi India (Govt of NCT Delhi, India). He has published **Forty** research papers in referred International Journals. Dr Khandelwal is serving as a reviewer in various renowned journals i.e. IEEE MTTs (USA), IEEE Access (USA), IET Microwave Antenna and Propagation (UK), IET Electronics Letters (UK), Wireless Personal Communication (Springer, Netherland), International Journal of Electronics and Communications (Elsevier), Journal of Electromagnetic Waves and Applications, Taylor & Francis (UK) etc. His current research interests focus on microstrip antennas, metamaterials, 5-G antennas, UWB antennas, rectennas, reconfigurable antennas, multiband antennas, circularly polarized antennas, antenna array, defected ground structure and microwave components.

Trap Assisted Persistent Photo-Conductivity in Solution-Processed CuO Thin Film

Akash Tripathi, Jitesh Agrawal, Tejendra Dixit, and Vipul Singh

***Abstract*— Optically induced meta-stable states in the semiconductor lattice temporarily change the conductivity which leads to a phenomenon called persistent photo-conductivity (PPC). We report a PPC effect with long retention time of more than 90 h in solution-processed highly transparent CuO thin film. A 254 nm wavelength light source was used to 'set' and thermal annealing at 50°C for 10 minutes and NIR illumination was used to reset PPC effect. Activated trap states without any stimulus which are responsible for PPC effect in CuO thin film has, been investigated by performing temperature-dependent I-V characterization. These trapped charges are released either on heating or illumination with NIR light source. Such a long-enduring PPC achieved in low-cost transparent CuO thin film could pave the way for further study of the feasibility of optically defined, transparent electronics.**

***Index Terms*— CuO, persistent photo-conductivity, solution-process, thin film device, trap states.**

I. Introduction

THE study of interaction of photon and matter is one of the broad fields of science. Photoinduced conductivity in the semiconductors is one such field which has been intensively investigated by the scientific community for many decades. The phenomena of PPC was first defined by Henry and Lang [1] at Bell Laboratories as large photo-conductivity which persists for hours or even days, after the light is switched off. As discussed by Lany and Zunger defect localized states (DX centers) [2] act as the trap state for the free charge carriers and reduces the recombination rate which results in enhanced lifetime of the optically generated electron and hole pairs. These trap states could increase the carrier lifetime to several days which leads to the origin of PPC [3]. This state of meta-stability in conductivity was early on reported at low temperatures and only phenomenological study was performed [1], [4]–[7]. To utilize this physical phenomenon towards real-life applications there are several reports on room temperature PPC [8], [9].

Initially, PPC was considered as a drawback for the functioning of the photo-detector due to its sluggish response to light [9]. On the other hand this meta-stability in conductivity has been utilized for applications such as holographic data storage, writing low resistance path between two contacts, optical memory and high-speed infrared photo-detector [6], [10]–[12]. For utilizing PPC in above-mentioned applications 'reset' mechanism was also developed viz. thermal annealing and NIR quenching [6], [9], [12].

PPC has been demonstrated in semiconductors like $SrTiO_3$, GaN, GaAs which are costly rare earth materials. On the other hand metal oxides can be a chipper alternative. CuO is metal oxide semiconductor with an indirect bandgap of 1 to 1.4 eV and the direct bandgap of 2 to 4 eV [13]. Solution based growth process like hydrothermal technique facilitates facile, large area and cost-effective synthesis of CuO [14], [15]. Another advantage of CuO is that it is a p-type semiconductor which is crucial to realize all oxide devices like PN junction diode and CMOS transistors [15], [16]. It has been utilized in optoelectronic applications like photovoltaic solar cell, photo-detectors [15], [17]–[19].

We report a highly endurable PPC in solution-processed CuO thin film at room temperature which lasted for several days. As synthesized CuO thin film was highly transparent with more than 90% transparency in the visible region. The bandgap was calculated to be 3.67 eV and hence 254 nm wavelength light source was suitable to optically induce meta-stable state which results in PPC. We successfully demonstrate suppression of PPC effect by thermal annealing at 50°C for 10 minutes and optical quenching by NIR (800 nm) light. To investigate the PPC phenomena in CuO thin-film temperature-dependent I-V was performed. After investigation it was found that active trap states were present in the pristine CuO sample. After illuminating the CuO thin film with 254 nm light source there was formation of defect states near the conduction band edge. These additional optically induced defect states could also be responsible for such a large PPC. Trap states assisted PPC achieved that to in the low-cost transparent CuO synthesized by a facile method could pave the way for further study of the feasibility of optically defined, transparent electronics.

II. Experimental Section

Three steps cleaning process was incorporated for cleaning the glass substrate viz: ultra-sonicating in acetone, isopropanol and de-ionized (DI) water, respectively for 10 minutes each. Consecutively hydrophilic treatment of all samples was performed in the solution of H_2O:NH_4OH:H_2O_2 (volume ratio 5:1:1) for 20 minutes at 60 °C. For deposition of CuO thin film over the substrate by solution process, seed

Manuscript received August 1, 2019; revised September 24, 2019; accepted October 25, 2019. Date of publication November 8, 2019; date of current version November 27, 2019. This work was supported by the Council of Scientific and Industrial Research for Fellowship under Grant 09/1022(0045/2017-EMR-1). *(Corresponding author: Vipul Singh.)*

A. Tripathi, J. Agrawal, and V. Singh are with the Molecular and Nanoelectronic Research Group, Department of Electrical Engineering, IIT Indore, Indore 453552, India (e-mail: vipul@iiti.ac.in).

T. Dixit is with the Nano Functional Materials Technology Centre, IIT Madras, Chennai 600036, India, with the Department of Physics, IIT Madras, Chennai 600036, India, and also with the Materials Science Research Centre, IIT Madras, Chennai 600036, India.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JQE.2019.2952385

0018-9197 © 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

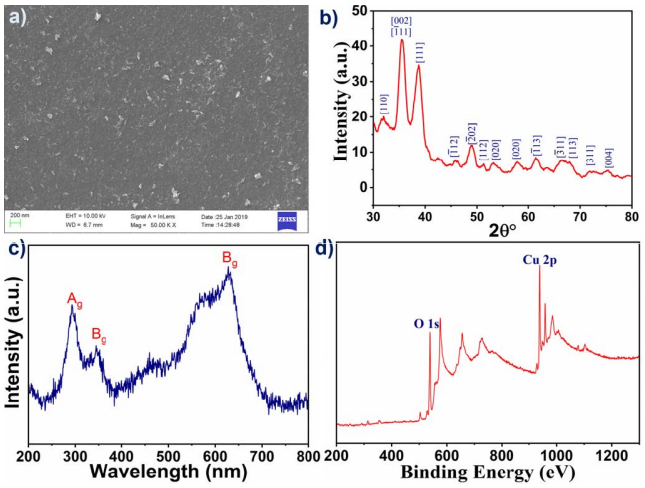


Fig. 1. (a) FESEM image, (b) GIXRD pattern, (c) Raman spectra and (d) XPS spectra of as synthesized CuO thin-film.

solution was prepared by continuous stirring of copper acetate monohydrate ($Cu(CO_2CH_3)_2H_2O$) in DI water for 4 hrs. Seed solution was spin-coated over the glass substrate at 1500 rpm for 30 sec. The obtained sample was annealed for 30 minutes at 200°C. DI water was used to rinse the obtained sample to remove the impurities and a layer of uniform thin film of CuO was left behind.

To investigate the surface of the film field emission scanning electron microscopy (FESEM), Zeiss Supra-55 was used. The crystalline property was investigated by XRD (GIXRD, Rigaku SmartLab®system) with Cu-Kα radiation ($\lambda = 1.54$ Å) source. Raman spectroscopy was done using Horiba, Labram HR evolution, from 200 to 800 cm^{-1}. X-ray photoelectron spectroscopy (XPS) was performed by Axis-Ultra DLD system. To study the absorbance of the film UV-Vis spectroscopy was done in the range 200 nm to 800 nm on Cary 60 UV-Vis, Agilent Technologies.

Gold (Au) electrodes were deposited by physical vapour deposition technique at the rate of 1 Å/s. The thickness of the electrodes was 50 nm and the gap between the electrodes was 50 μm. Keithley 2612A dual-channel source meter was used for electrical characterization. UV-C light was used as the optical excitation source to carry out photoconductivity measurements. Four probe measurement was done to measure the sheet resistance.

III. Results and Discussion

FESEM image of the thin film as shown in Fig. 1(a) confirms the uniformity of the film. The thickness of the film was observed to be in the order of 50 nm. The phase purity of the film was confirmed by performing Grazing incidence X-ray diffraction (GIXRD) at an incidence angle of 0.3°C. The GIXRD data was compared with standard JCPDS data [15] and all the peaks correspond to the monoclinic crystal structure of CuO (Fig. 1(b)). The Scherrer equation used to calculate the mean crystallite size was given as:

$$D = \frac{k\lambda}{\beta \cos\theta}$$

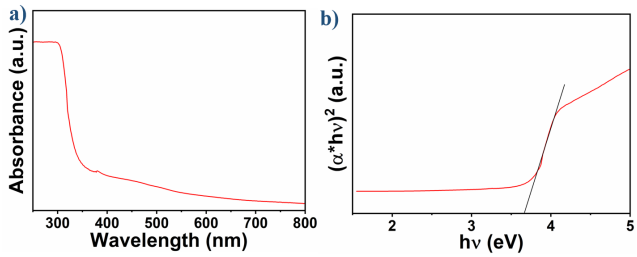


Fig. 2. (a) Absorbance and (b) Tauc plot of the as synthesized CuO thin-film.

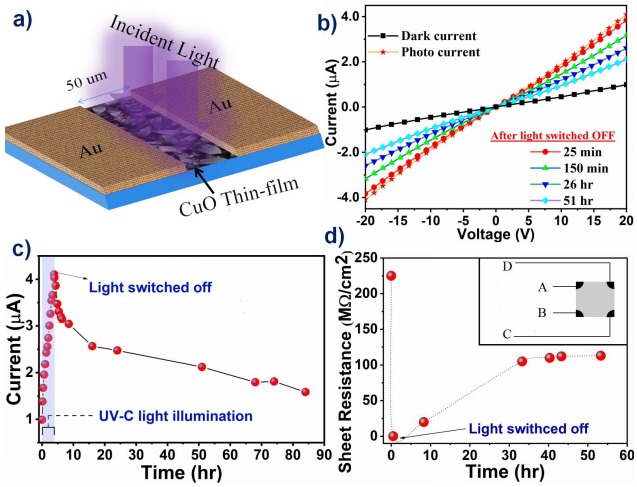


Fig. 3. (a) Device schematic, (b) I-V characteristic of the device in dark, illuminated and after switching off the 254 nm light source, (c) current value with respect to time at a fixed bias of 20 V and (d) sheet resistance variation with respect to time and inset shows schematic representation the arrangement for resistivity measurement.

where D is the mean crystallite size, k is the shape factor, λ is the X-ray wavelength, β is line broadening at half the maximum intensity and θ is the Bragg angle. The calculated average crystallite size was ~10 nm.

Raman spectra of the thin film in Fig. 1(c) has shown three different peaks. All these peaks belong to 3 Raman active modes (A_g+2B_g). The peak at 300 cm^{-1} belongs to A_g and the peak at 323 cm^{-1} and 640 cm^{-1} belongs to B_g mode of CuO. XPS analysis of the CuO film was performed to further investigate the phase purity of the sample. Fig. 1(d). shows XPS survey scan and all the peaks well matched with the pure CuO [20].

To investigate the optical property of the as-synthesized CuO film UV-Vis spectroscopy was performed (Fig. 2(a)). The absorption edge was found to be near at 300 nm and the film was highly transparent in the visible region. Fig. 2(b) shows the Tauc plot, with the bandgap of the as-synthesized CuO film is 3.67 eV. The as-synthesized CuO thin film could find novel application in transparent photovoltaic windows.

A planner device with top electrode configuration was fabricated (device schematic is shown in Fig. 3(a)) as specified in the experimental section. The I-V characterization of the device in dark condition has been shown in Fig. 3(a). At 20 V applied bias the dark current value was 1 μA.

The dark current curve was linear in nature which indicated the formation of ohmic contact between the CuO (Φ_{CuO} = 5.3 eV) and Au (Φ_{Au} = 5.1 eV) electrodes. To perform optical characterization, the device was illuminated with a light source of wavelengths ranging from 254 nm to 800 nm. It was observed that as-synthesized CuO thin film-based device is almost insensitive towards the wavelength of more than 290 nm. In convergence with the absorbance data the device has shown solar-blind behavior which can be utilized for many other optoelectronic applications viz. flame detection, missile tracking systems, etc. Upon exposing the device with 254 nm light source (incident power of 5 mWcm^{-2}), the photo-current saturates to value of 4.05 μA. As the light was switched off, rather than instantaneous suppression, there is very gradual decrease in photo-current (Fig. 3(c)). The residual photo-current persists for 90 h (3.75 days) and it takes even more time to suppress completely which could be assigned to the formation of metastable state.

The observed persistent photo-conductivity was further investigated by measuring the sheet resistance of the CuO thin film by using van der Pauw method [21]. Four contact point A, B, C and D were made at the four corners of the square CuO thin film sample (1.5 cm*1.5 cm) (inset of Fig. 3(d)). After illuminating the sample with 254 nm wavelength light, the sheet resistance of CuO thin film decreases by 2 orders (Fig. 3(d)). As the light is switched off again the semiconductor does not attains it's original sheet resistance value till the measurement was last performed (54 h).

In order to make use of PPC phenomenon, both Set and Reset mechanism are required. As we have already discussed, the device was 'Set' by inducing the metastable state in the CuO thin film via UV-C illumination. Now to make this effect utilizable a 'reset' mechanism has been established. The device was reset to its high resistivity state by thermal annealing. After annealing the device at 50°C for 10 minutes, the PPC effect was observed to be completely suppressed and device recovered to its original conduction state.

Another possible way to optically reset the device is to illuminate the sample with NIR light. Therefore, in order to reset the device which was already in set condition, has been illuminated with 800 nm light source. As shown in Fig. 4(a), after NIR light is switched On, the photo-current start to decrease rapidly. Interestingly, the current continues to decrease even after the NIR light is switched off which suggests that NIR illumination has triggered the reset mechanism which was responsible for the continuous decrease in the photo-current. In comparison to the PPC retention time of more than 90 hours, it took only 200 minutes to reset the device using this method. This optical reset mechanism achieved in transparent CuO thin film can be further utilized in optically defined transparent electronics applications.

The optical reset mechanism was performed at room temperature which indicates that PPC in CuO is due to random local potential fluctuations induced by defect states [6], which indicates that there are defect states present in the as-synthesized CuO. Additionally, changing the semiconductor's conductivity by light illumination and getting back to its pristine conduction state by thermal annealing is clear evidence that there is

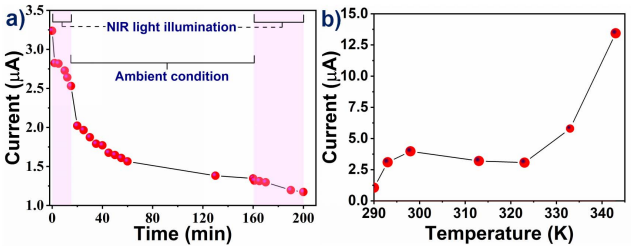


Fig. 4. Value of current (V_B = 20 V) with respect to (a) time during the optical reset process by NIR (800 nm) light source and (b) temperature.

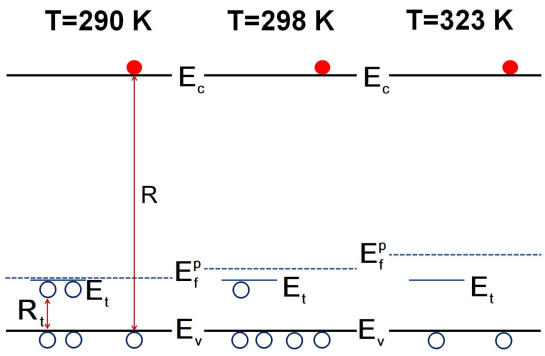


Fig. 5. Band diagram of CuO showing shift in quasi-Fermi level with respect to temperature.

a presence of trap levels or sub-band levels that holds the charge carriers and this process contributes to the conduction mechanism [3].

Notably, Bhatnagar *et. al.* have probed the trap levels by the measurement of thermally stimulated currents (TSC) [3]. To probe trap states in the CuO thin-film temperature-dependent I-V characterization of the device was performed in the pristine state with no prior optical excitation (Fig. 4(b)). A bias voltage (V_B) of 20 V was applied across the electrodes and temperature was varied from 290 k to 343 k. With the increase in temperature the current first increases from 1 μA at 290 K to 4 μA at 298 K. As the temperature is increased above 298 K the current starts to decrease and beyond 323 K it again starts to increase.

This behavior of current variation in pristine CuO can be explained by assuming that the charge carriers are already in trap state without any external stimulus. As shown in Fig. 5 at temperature T=290 K, the quasi-Fermi level of holes (E_f^p) is very near and just above of the trap state (E_t). Therefore, the trapped charge carriers would be in thermal equilibrium with the valence band. Consequently, the rate of hole transfer (R_t) from E_t to valence band is more as compared to rate of charge carrier recombination (R). As the temperature increases, the quasi-Fermi level E_f^p gradually shifts towards the mid of the bandgap. The peak current of 4 μA was observed at T=298 K, which was assigned to maximum hole concentration in the valance band. Beyond this temperature, R overtakes R_t and the concentration of holes in valence band starts to decrease hence the current also decreases. Further, as the temperature is increased beyond 323 K, the thermal

generation of the rate of electron and hole pairs supersede R and hence the conductivity is enhanced with the temperature.

From the above discussion it is inferred that CuO has trap levels which are active at room temperature and these trap states are capable of holding the charge carrier for long duration. Hence, as this device is illuminated with the light source of photon energy more than the bandgap i.e. 3.67 eV the charge carriers will be trapped which resulted in decreased rate of recombination due to which carrier lifetime is also increased. Therefore, the photo-conductivity persists for a very long duration.

Transmittance analysis of the pristine and illuminated sample of CuO thin film was also performed to investigate the effect of UV-C exposure on the optical property of the sample (as discussed in our previous work [15]). It was observed that after UV-C illumination, CuO thin film transmittance is decreased in the range 400 nm to 300 nm which suggests that there is formation of defect states near the conduction band edge. These defect states which were created after illumination of the sample could also behave as trap states apart from the already present trap states. This will be probed in our future work.

IV. Conclusion

Solution-processed highly transparent CuO thin film is synthesized over the glass substrate. A planar device with top Au electrodes is fabricated with solar-blind photo-detection characteristics. The device was illuminated with 254 nm light source and a very large PPC was observed which lasted for more than 90 h. Two reset mechanisms were demonstrated viz. thermal annealing and NIR light illumination. NIR light was able to suppress the PPC effect after 200 minutes of illumination. The trap states which are active at room temperature are responsible for PPC. Set and reset mechanism can be utilized in future application like optical memory and making reconfigurable contacts optically between the communicating devices.

Acknowledgment

The authors would like to acknowledge SIC, IIT Indore for providing the research facilities.

References

[1] C. H. Henry and D. V. Lang, "Nonradiative capture and recombination by multiphonon emission in GaAs and GaP," *Phys. Rev. B, Condens. Matter*, vol. 15, no. 2, pp. 989–1016, Jan. 1977.

[2] S. Lany and A. Zunger, "Anion vacancies as a source of persistent photoconductivity in II-VI and chalcopyrite semiconductors," *Phys. Rev. B, Condens. Matter*, vol. 72, Jul. 2015, Art. no. 035215.

[3] A. Bhatnagar, Y. H. Kim, D. Hesse, and M. Alexe, "Persistent photoconductivity in strained epitaxial $BiFeO_3$ thin films," *Nano Lett.*, vol. 14, no. 9, pp. 5224–5228, 2014.

[4] M. G. Craford, G. E. Stillman, J. A. Rossi, and N. Holonyak, Jr., "Effect of Te and S donor levels on the properties of $GaAs_{1-x}P_x$ near the direct-indirect transition," *Phys. Rev. J. Arch. Published Amer. Phys. Soc.*, vol. 168, no. 3, pp. 867–881, 1968.

[5] J. Klem *et al.*, "Persistent photoconductivity in (Al,Ga)As/GaAs modulation doped structures: Dependence on structure and growth temperature," *J. Appl. Phys.*, vol. 54, no. 9, p. 5214, 1983.

[6] H. X. Jiang, G. Brown, and J. Y. Lin, "Persistent photoconductivity in II-VI and III-V semiconductor alloys and a novel infrared detector," *J. Appl. Phys.*, vol. 69, no. 9, pp. 6701–6703, 1991.

[7] P. Bhattacharya, J. Ku, S. Oevn, G. Olsen, and S.-H. Chiao, "LPE and VPE In1-xGaxAsyP1-y/lnP: Transport properties, defects, and device considerations," *IEEE J. Quantum Electron.*, vol. 17, no. 2, pp. 150–161, Feb. 1981.

[8] C. Stockinger, W. Markowitsch, W. Lang, W. Kula, and R. Sobolewski, "Mechanisms of photodoping in oxygen-deficient $YBa_2Cu_3O_x$ films studied by *in situ* transport measurements," *Phys. Rev. B, Condens. Matter*, vol. 57, no. 14, pp. 8702–8708, Apr. 1998.

[9] M. Hou, H. So, A. J. Suria, A. S. Yalamarthy, and D. G. Senesky, "Suppression of persistent photoconductivity in AlGaN/GaN ultraviolet photodetectors using *in situ* heating," *IEEE Electron Device Lett.*, vol. 38, no. 1, pp. 56–59, Jan. 2017.

[10] L. Hesselink, S. S. Orlov, and M. C. Bashaw, "Holographic data storage systems," *Proc. IEEE*, vol. 92, no. 8, pp. 1231–1280, Aug. 2014.

[11] V. M. Poole, S. J. Jokela, and M. D. McCluskey, "Using persistent photoconductivity to write a low-resistance path in $SrTiO_3$," *Sci. Rep.*, vol. 7, Jul. 2017, Art. no. 6659.

[12] L. X. He, K. P. Martin, and R. J. Higgins, "Infrared quenching of persistent photoconductivity in $GaAs/Al_xGa_{1-x}As$ heterostructures," *Phys. Rev. B, Condens. Matter*, vol. 39, no. 3, pp. 1808–1818, Jan. 1989.

[13] Q. Zhang *et al.*, "CuO nanostructures: Synthesis, characterization, growth mechanisms, fundamental properties, and applications," *Prog. Mater. Sci.*, vol. 60, pp. 208–337, Mar. 2014.

[14] S. Lee *et al.*, "Sol-gel processed p-type CuO phototransistor for a near-infrared sensor," *IEEE Electron Device Lett.*, vol. 39, no. 1, pp. 47–50, Jan. 2018.

[15] T. Dixit, A. Tripathi, K. L. Ganapathi, I. A. Palani, M. S. R. Rao, and V. Singh, "Solution-processed transparent CuO thin films for solar-blind photodetection," *IEEE Electron Device Lett.*, vol. 40, no. 2, pp. 255–258, Feb. 2019.

[16] K.-C. Wang *et al.*, "p-type mesoscopic nickel oxide/organometallic perovskite heterojunction solar cells," *Sci. Rep.*, vol. 4, Apr. 2014, Art. no. 4756.

[17] P. Wang, X. Zhao, and B. Li, "ZnO-coated CuO nanowire arrays: Fabrications, optoelectronic properties, and photovoltaic applications," *Opt. Express*, vol. 19, no. 12, pp. 11271–11279, May 2011.

[18] P. Raksa, S. Nilphai, A. Gardchareon, and S. Choopun, "Copper oxide thin film and nanowire as a barrier in ZnO dye-sensitized solar cells," *Thin Solid Films*, vol. 517, pp. 4741–4744, Jul. 2009.

[19] H. J. Song *et al.*, "High-performance copper oxide visible-light photodetector via grain-structure model," *Sci. Rep.*, vol. 9, May 2019, Art. no. 7334.

[20] T. S. Jung *et al.*, "Enhancement of switching characteristic for p-type oxide semiconductors using hypochlorous acid," *ACS Appl. Mater. Interfaces*, vol. 10, no. 38, pp. 32337–32343, 2018.

[21] S. H. N. Lim, D. R. McKenzie, and M. M. M. Bilek, "van der Pauw method for measuring resistivity of a plane sample with distant boundaries," *Rev. Sci. Instrum.*, vol. 80, Jul. 2009, Art. no. 075109.



Akash Tripathi was born in Azamgarh, India, in August 1989. He received the B.Tech. degree in electronics and communication engineering (ECE) from the Invertis Institute of Technology, Bareilly, India, in 2014, and the M.Tech. degree in VLSI and embedded system design from the Maulana Azad National Institute of Technology, Bhopal, India. He is currently pursuing the Ph.D. degree with the Indian Institute of Technology Indore, India. He worked on the SMDP C2SD project to Design an ASIC for Encryption and Decryption of Serial Data for Transceiver from July 2016 to June 2017. He is also working on as and holds a Senior Research Fellowship (SRF) from the Council of Scientific and Industrial Research.



Jitesh Agrawal was born in Nagda, India, in August 1989. He received the B.E. degree in electronics and communication engineering from the Mahakal Institute of Technology, Ujjain, India, in 2010, and the M.E. degree from the Shri Vaishnav Institute of Technology, Indore, India, in 2013. He is currently pursuing the Ph.D. degree from the Indian Institute of Technology Indore, India. He is also working on metal oxide-based photodetector and holds a Senior Research Fellowship (SRF) from University Grants Commission, New Delhi, India.



Tejendra Dixit was born in Gorakhpur, India. He received the B.Tech. degree in electronics and communication engineering from the Ajay Kumar Garg Engineering College, Ghaziabad, India, in 2012, and the Ph.D. degree from the Indian Institute of Technology Indore, India. He has worked on metal oxide-based photodetector and also studied plasmonic behavior of noble metal nanoparticles. He is currently a Post-Doctoral Fellow with IIT Madras, in the area of 2-D Semiconductors. He has around 32 research publications in peer-reviewed journals.



Vipul Singh was born in Siwan, Bihar, India, in 1979. He received the bachelor's degree in electronics and communication engineering from the Delhi College of Engineering, University of Delhi, in 2003, and the M.S. and Ph.D. degrees from the Kyushu Institute of Technology, Japan, in March 2009. He did his Ph.D. research work with the Department of Biological Functions and Engineering, Kyushu Institute of Technology, under the supervision of Prof. K. Kaneto. Most of his Ph.D. work was based on study of interface related effects in conducting polymer-based organic optoelectronic devices. From April 2009 to November 2010, he was a Scientific Researcher with the Nanosystems Integration Laboratory, Research Institute of Electronics, Shizuoka University, Japan, where he studied MOSFET-based Single electron Transistors/ devices. Since then, he has been interested in Bulk/SOI technologies and has also studied and reduced low-frequency noise occurring in these devices due to interface traps both at room and cryogenic temperatures. Since November 2010, he has also been working as an Assistant Professor with the Department of Electrical Engineering, Indian Institute of Technology (IIT), Indore. His research interest are in the area of organic electronics ranges from device applications, namely, organic field effect transistors (OFETs), organic solar cells (OSCs), and organic light emitting diodes (OLEDs) to the study of charge carrier transport or photo carrier generation in various organic materials.

International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 07 | July 2019 www.irjet.net p-ISSN: 2395-0072

Water Footprint Analysis of Delhi to Understand it's sustainability

Vaibhav[1], Dr Bharat Jhamnani[2]

[1]*Master of technology scholar, Department of Environmental Engineering, Delhi Technological University, Shahbad Daulatpura, Main Bawana Road, Delhi, 110042, India*
[2]*Assistant professor, Department of Civil Engineering, Delhi Technological University, Shahbad Daulatpura, Main Bawana Road, Delhi, 110042, India*

Abstract - *India's ambition to meet SDGs (Sustainable Development Goals) like zero hunger and access to clean water depends upon the strategic use and management of limited freshwater resources. To do so, water consumption in various sectors must be established rather than using the traditional water withdrawal method. Water footprint (WF) is an indicator used to find the volume of fresh water being used in product, process, by a consumer, community, or nation directly and indirectly. It is based on the concepts of material flow analysis (MFA) and life cycle assessment (LCA). It traces all the processes and the direct water being consumed in them, giving virtual water content (VWC, water required to produce a unit product) of a particular product.*

In this study, it was found WF of Delhi in litres per capita per day (lpcd) for the year 2011 and forecasted for the year 2021 using water footprint assessment (WFA). The average water footprint of a person living in Delhi in the year 2011 came out to be 2913 lpcd broken down as crop WF: 1868 lpcd; livestock WF: 561 pcd; miscellaneous (comprises of products necessary in urban living such as petrol, liquor, etc.) WF: 207 lpcd and direct WF: 277 lpcd. It was found in the year 2021; crop WF is going to increase marginally because of higher dairy products and meat consumption along with lower consumption of cereals. Increase of 12.8% in total WF from 2011 to 2021 is primarily because of direct WF. There are many projects underway by the Delhi Jal Board (DJB) such as Renuka Dam, Kishau Dam and Lakhwar Vyasi Dam which aim to increase direct water supply by 66% i.e. 258 lpcd in 2011 to 430 lpcd in 2021.

Interestingly ~85-90% of WF of a person in Delhi is due to indirect consumption in the form of crop and livestock products which are being imported since there are no agricultural practices in Delhi. Remaining ~10-15% is in the form of direct WF, most of it comes from places which are not in geographical boundaries of Delhi (although they are a part of Delhi Hydrometric Area (DHA)). Delhi is almost entirely dependent on direct and indirect WF even though Delhi receives ~ 1 m of average rainfall every year.

***Key Words*: Environmental footprint; Water footprint; Water footprint assessment; Virtual water**

1. INTRODUCTION

Cities are the congested economic hubs of any nation; it is likely that many cities will face challenges in meeting the increasing demands of land, energy, water and other limited resources (UN, 2018). Freshwater use in agricultural, industrial, and household processes is increasing, which makes sustainable water use and management irrefutable. Stress on limited freshwater water resources is rapidly increasing in cities due to increasing population density, changing lifestyle, and urbanization (Gleick, 2000). The total urban population in India grew from 17.97% in 1961 to 31.16% in 2011 (Bhagat, 2011).

Freshwater resources are unevenly distributed all around India and the world with some areas being water scarce, while some are abundant. However, through the trade of goods and services, freshwater is being exchanged inter-regionally and internationally. This is known as virtual water (or indirect water) trade (Allan, 1996). Virtual water became a building block of an indicator known as "Water Footprint" (WF) developed by Hoekstra and Hung in 2002. Water footprint network (URL 1) has been developed over the years with a dedicated group of researchers working on water footprint. The water footprint is used to account actual or true water consumption for a product, process, consumer, community, business, or a nation (Hoekstra, Chapagain, & Aldaya, 2011).

1.1 Classification of Water Footprint

There are two classifications of water footprint. First, it is classified as direct and indirect to distinguish between the natures of the WF. Direct WF is water consumed directly, such as drinking, washing, etc. and indirect WF is water consumed indirectly through the purchase of goods or services such as buying foods, clothes, etc. Second, it is divided into blue, green, and grey WF to account for various water sources and water quality levels. The blue WF refers to the water taken up from groundwater and surface water resources during the production of a commodity. The green WF is an indicator of the precipitated water that does not run off or recharge the groundwater but gets stored as moisture in the soil and ultimately taken up by plants for their growth or gets evaporated. The grey WF is the water required to dilute the water, coming out from a production line of a product, according to water quality standards

(Mekonnen & Hoekstra, 2011). It is important to understand the difference between the water footprint of production (WF_{prod}) and the water footprint of consumption (WF_{cons}). The former is the sum of direct and indirect water use of regional water resource used in the process of production and latter is the sum of direct and indirect water use of regional and foreign water resource through domestic consumption (Vanham, Hoekstra, & Bidoglio, 2013).

1.2 Assessment Techniques

WF traces the flow of direct water consumed in every process involved. WF uses the concepts of material flow analysis (MFA) and life cycle assessment (LCA). However, WF measures on one attribute/unit, i.e., the volume of water and do not measure impact (Hoekstra et al., 2011). Common environmental impacts associated with excessive freshwater use are groundwater depletion, biodiversity loss, and freshwater pollution (Gleick, 2006).

1.3 Water Footprint of a city

WF studies now are being more focused on the city scale, such as Zhang et al., 2011, Feng et al., 2010 and (Ahams et al., 2017). Studies at city level give information about the allocation of water resources at a local scale. This allows taking appropriate measures and changes to optimize freshwater use. Cities are a good market place to sell goods which are not locally produced. Thus, cities use local water sources mostly for direct water consumption while putting burden and stress for indirect consumption on water resources somewhere else where they are being manufactured. Water saving in cities can be created by trading from a place which requires lesser Virtual Water Content (VWC, water required to produce or manufacture a unit of a particular product) than from a place having higher VWC for the same product (A. K. Chapagain, Hoekstra, & Savenije, 2006).

1.4 Scope of the study

In this study, we applied WFA (Water Footprint Assessment) to find per capita consumption of water in Delhi as direct and indirect WF. WF of consumption can then help to find which sector or activities have higher water utilization and how can it be reduced.

In India's fight against poverty and hunger and to achieve food and water security, WF can play a huge role in the future, particularly in the agricultural sector (URL 2).

Around 80% of fresh water is being used in agriculture (K Bharat & B Dkhar, 2018). Both food and water security are interconnected, and they should be looked with a more holistic approach at a local level. Current water productivity in agriculture in India is lower than that of the global average. With rapid population growth, it seems inevitable for India to become a water scarce nation (Rockstrom, Lannerstad, & Falkenmark, 2007). To tackle the situation, some possible ways are like increasing crop productivity, changing consumption patterns, or decreasing wastes and losses of agricultural products, all of which must go hand in hand.

Similar to other environmental indicators, water footprint does not give a complete picture of every environmental aspect. Therefore, it should be integrated with other relevant indicators to get social, economic, and environmental insights. The WF can act as a partial tool in the decision-making process (Hoekstra et al., 2011).

2. METHODOLOGY AND DATASETS

2.1 Method

This section describes the approach followed to determine and forecast the total water footprint in litres per capita per day (lpcd) and the losses as virtual water in different agricultural products for Delhi for the years 2011 and 2021.

2.2 Water Footprint Calculation

The summation of the footprint of the individual product gives the total water footprint within the geographical boundaries of Delhi as given by equations below.
where,

$$WF_{Delhi}=WF_{direct}+WF_{indirect} \quad [1]$$

WF_{direct} = Direct WF = Water being supplied by DJB (Delhi Jal Board) and groundwater used by households and industries in Delhi

$WF_{indirect}$ = Indirect WF = Water being used indirectly in products such as food, clothes, petrol etc.

$$WF_{direct} = \sum_{i=1}^{n} DW_i \quad [2]$$

where,
DW_i= Direct water through i^{th} source
i = 1, 2….n

$$WF_{indirect} = \sum_{p=1}^{m} WF_p \quad [3]$$

$WF_p = VWC_p * w$
where,
WF_p = Water footprint of p^{th} product
p = 1, 2…m
VWC_p = Virtual Water Content of product p (m^3/ton or l/kg)
w = weight of product p consumed (ton or kg)

2.3 Forecast

The trend of consumption of crop and livestock products for Delhi is assumed to be the same as for average India. The trend of consumption of various products was found out by

using the FAO database. Also, virtual water content for any product p (VWC_p) is estimated to remain constant temporarily.

Time series curves were developed for each agricultural product from the data collected from FAO (as shown in Appendix figures). Outliers were identified by using box plot of residuals.

The equation used to find consumption for the year 2021:

$$C_{2021,p} = \frac{C, NSSO_{2011,p}}{C, FAO_{2011,p}} \times C, Estimated\ FAO_{2021,p}$$

where,
$C_{2021,p}$: Consumption for product p (Delhi 2021)
$C,NSSO_{2011,p}$: Consumption for product p (Delhi, 2011) (Source: NSSO, 2014)
C, FAO2011, p: Consumption for product p (Indian average, 2011) (Source: URL C)
C, Estimated, FAO2021, p: Consumption for product p (Indian average, 2021).
It was found out by using the best fit trend line for product p using FAO data from 1961 to 2013 (Source: URL C).
Note: While forecasting, we excluded miscellaneous WF from our system boundary.

2.4 Virtual Water Content (VWC) of Various Crops and Products

Virtual water content (VWC) values were majorly taken from various Water Footprint Network (WFN) sources.

2.5 Datasets

All the data collected to calculate the water footprint of Delhi were for the year 2011-12. The main reason being, the last census report was published in 2011 and hence, better accuracy on per capita consumption. Data was collected from different sources, suitable assumptions were made, and it has been reported in detail in the supplementary material. No primary data and industrial data were collected. The system boundary for this study is shown in figure 1.

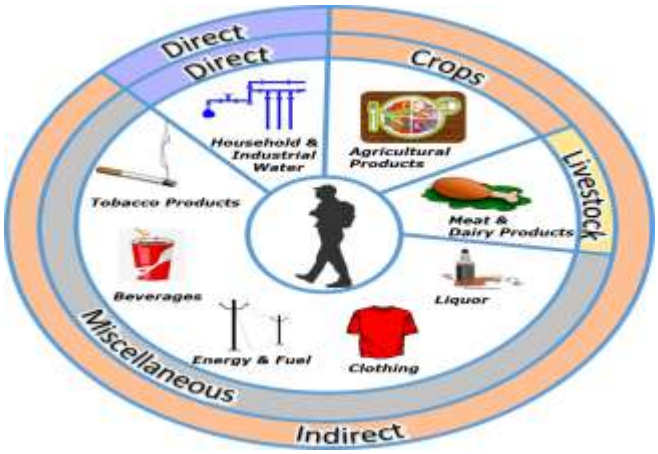


Fig -1: System boundaries for this study

2.5.1Direct Water Use and its Footprint

Direct water consumed in Delhi comprises of surface water from various rivers and groundwater.

Total direct water footprint in Delhi in 2011:

WF_{direct}= 1020 MGD (Surface and Ground Water Use) + 0.34 MGD (Virtual water associated with the treatment of water, Neglected as it less than 0.03% of direct water use)
= 1020 MGD
= 3861MLD
= 277 litres per day per capita
(Population of Delhi in 2011 = 16.7 million)

2.5.2Indirect Water Use and its Footprint

A total of 104 products were considered in this report to find the water footprint of Delhi. 75 products in crops, 13 products in livestock, and 16 miscellaneous products, which are important in urban living were included in this study.
Agricultural Products considered are as follows:

Rice – P.D.S., Rice – other sources, Chira/Flattened rice, Khoi, lawa(puffed rice), Muri(puffed rice), Other rice products, Wheat/atta – P.D.S., Wheat/atta – other sources, Maida/wheat flour, Bread: bakery, Other wheat products, Jowar(sorghum) and its products, Bajra(pearl millet) and its products, Maize and its products, Barley and its products, Small millets and its products, Ragi(African millet) and its products, Other cereals, Arhar/Pigeon pea , tur, Gram: split, Gram: whole, Moong/Green Gram, Masur/Red Lentil, Urd/Black lentils, Other pulses, Gram products, Besan or gram flour or chickpeas flour, Other pulse products, Sugar – PDS, Sugar – other sources, Gur/brown sugar, Mustard oil, Groundnut oil, Coconut oil, Refined oil, Edible oil: others, Potato, Onion, Tomato, Brinjal, Carrot, Palak(spinach)/other, Green chillies, Cauliflower, Cabbage, Gourd, pumpkin, Peas, Beans, barbate, Lemon, Other vegetables, Banana, Watermelon, Pineapple, Coconut , Green coconut, Papaya, Mango, Apple, Grapes, Groundnut, Dates, Cashewnut, Walnut, Other nuts, Pears/nashpati, Berries, Ginger, Garlic, Dhania/Coriander, Turmeric , Black pepper , Dry chillies, Oilseeds, Other spices, Milk: liquid, Milk: powder, Curd,
Livestock Products:

Ghee/hard fat, Butter, Cheese(kg), Eggs (no.), Fish, prawn, Goat meat/mutton, Beef/ buffalo meat, Pork, Chicken, Others: birds, crab, etc.,

Miscellaneous Products:
Tea cups (no.), Coffee cups (no.), Cold drinks (L), Fruit juice and shake (no.), Bidi(no.), Cigarettes(no.), Tobacco(gm), Liquor (litre), Electricity (std units), Kerosene (litre), Petrol (litre), Coal (kg), L.P.G. (kg), Leather footwear (pair), Total no. of clothes (T Shirt equivalent), Bed sheet (no.)

*All products were considered in kilograms unless mentioned otherwise.

2.6 Losses

The virtual water loss due to inefficiencies in production, processing, packaging, and distribution has been calculated using data from FAO for the South and Southeast Asia in the year 2007 to 2009. The rate of losses and waste was assumed to remain the same during the period 2011 to 2021.

3. RESULTS

3.1 Water Footprint of Delhi in 2011 and 2021

The average water footprint of a person living in Delhi in the year 2011 came out to be ~2913 lpcd with ~858 grams of average food consumption in a day. Detailed flow of WF from different products within system boundaries of this study is shown in figure 2. Food products (crop and livestock products) contribute ~ 85% of total WF.

It was found crop WF as 1868 lpcd, 64% of total WF. Consumption of cereals, pulses, and edible oil takes ~84% of crop WF. Cereals alone contribute to ~60% of crop WF. Consumption of rice products and wheat products has a 91.7% share in cereals and contribute to ~18% and ~21% of total indirect WF respectively.

WF of livestock products is 561 lpcd, which is ~19% of total WF. High consumption of milk (~166 grams/day) leads to high WF of dairy products. WF of meat per unit consumed is relatively high (~ 7000 l/kg) but lower consumption of meat results in lower WF, i.e. 192 lpcd.

Miscellaneous products take about 7% of total WF, i.e. 207 lpcd. WF of energy requirement in 2011 through petrol, kerosene, coal, and electricity is ~ 70 litres per person per day. This is lower than average WF of clothes bought, i.e. 106 lpcd.

Direct water received in Delhi in 2011 was about 277 lpcd through various rivers and groundwater sources.



Fig -2: Water footprint(lpcd) of Delhi in 2011

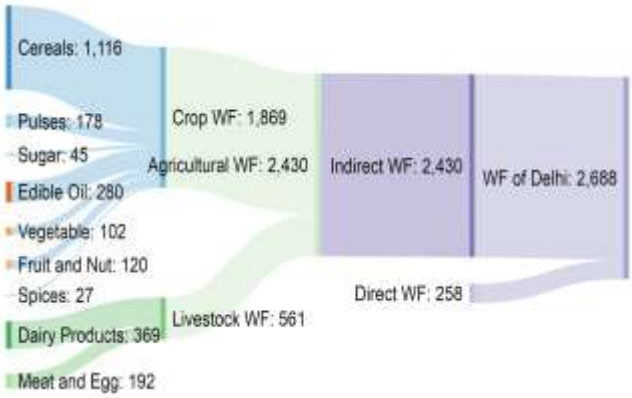


Fig- 3: Water footprint flow in Delhi 2011 in lpcd (excluding misc. products and groundwater)

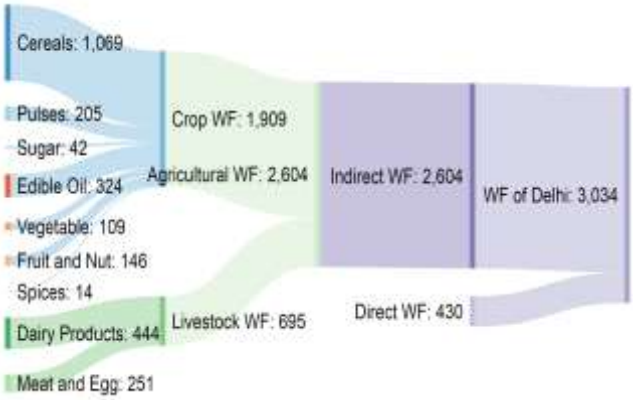


Fig- 0: Water footprint flow in Delhi 2021 in lpcd (excluding misc. products and groundwater)

The present study estimates WF of Delhi as ~ 3034 lpcd in 2021, based on projected input data of food products adapted from FAO database and neglecting miscellaneous products due to its insignificant contribution.

The results of this study have shown a 12.8% increase in total WF and ~6.5% increase in per day consumption from 858 grams per day in 2011 to 914 grams in 2021.

A ~4% decrease in WF was found in cereals, a product with the highest individual contribution. Spices and sugar consumption and WF has shown a marginal decrease. All other product categories have shown increased WF.

The WF of overall crop sector has been found to remain the same or increase marginally. Livestock WF is going to increase due to a shift in dietary patterns. Several projects are underway to increase direct WF by 66% in Delhi by DJB. Relative contribution of WF for year 2011 and 2021 is shown in figures 3 & 4. Comparison of the flow of WF through various product categories in the year 2011 and 2021 has been made in figure 5 and figure 6.

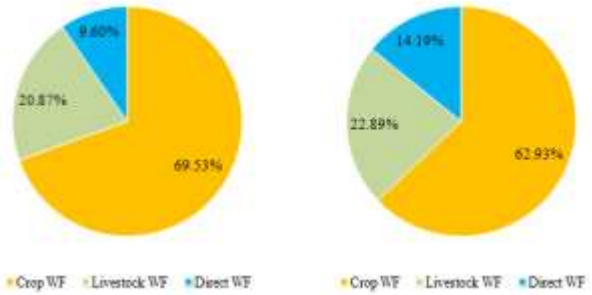


Fig -5: Sectoral water footprint in Delhi in (a)2011 and (b)2021

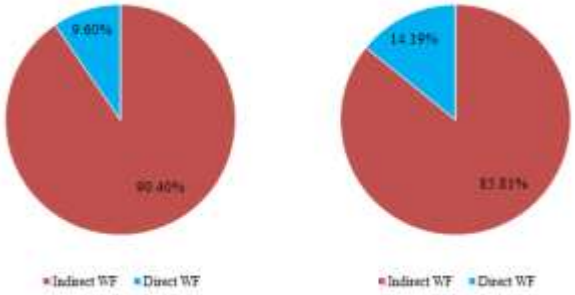


Fig -6: Distribution of types of WF in Delhi in (a) 2011 and (b) 2021

3.2 Losses

Due to inefficiencies in agricultural practices, the average virtual water lost in the year 2011 was 11239 million litres per day (MLD), which is going to increase to 12057 MLD in the year 2021. Average virtual water lost due to food wasted after consumption was 1243 MLD in the year 2011 and will become 1268 MLD in the year 2021. Average virtual water lost for different products has been tabulated in table 1.

Table -1: Losses of Virtual water in agricultural products (MLD) in 2011 and 2021 according to loss rates given by FAO

	Delhi 2011		Delhi 2021	
Agricultural Product	Virtual Water lost before consumption (MLD)	Virtual Water lost after consumption (MLD)	Virtual Water lost before consumption (MLD)	Virtual Water lost before consumption (MLD)
Cereals	3289	533	2845	461
Oilseeds & pulses	2377	82	2816	97
Fruits & vegetables	3275	389	3703	439
Meat	369	85	369	85
Fish & seafood	92	5	122	6
Milk & Cheese	1280	60	1539	72
Spices & Nuts	285	46	309	50
Sugar	134	22	156	25
Eggs	139	22	197	32
Total	11239	1243	12057	1268

3.3 Sensitivity Analysis

It was found that out of 38 categories of products, 5 products were contributing 73% of indirect WF, and the rest 33 products contributed the remaining 27%. Two scenarios were made to find sensitivity:

Scenario A: Variation in total WF due to variation in the water footprint of 33 products (contributing 27% of indirect WF) was studied. Other 5 products were assumed to have exact values of their WF.

Scenario B: Variation in total WF due to variation in the water footprint of 5 products (contributing 73 % of indirect WF) was studied. Rest 33 products were assumed to have exact values of their WF.

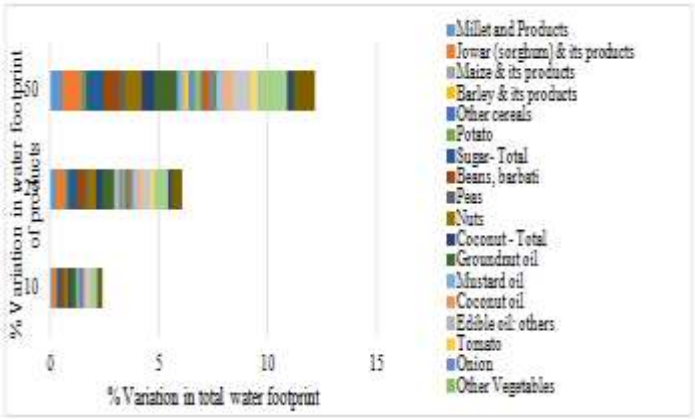


Chart -1: Percent variation in total WF with different percent variations in products according to scenario A

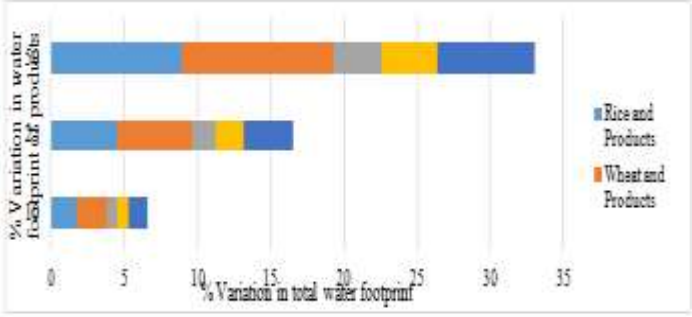


Chart -2: Percent variation in total WF with different percent variations in products according to scenario B

3.4 Validation

1. The data from NSS 68 of consumption of various agricultural products shows a correlation with the data given by FAO for the year 2011. The variation in data may be attributed to:
   - The difference in sampling methods by FAO and NSSO
   - FAO database uses January to December as a reference period of one year while NSSO conducted interviews from June'11 to July'12.
2. According to 'Water footprints of nations' by Chapagain & Hoekstra, 2004, average Indian internal agricultural and livestock WF of production = 2485 lpcd using analysis period as 1997–2001. This study gives agricultural and livestock WF of consumption in Delhi= 2429.2 lpcd in the year 2011-12.
3. According to 'The water footprint of India' by Kampman, 2007, the average Indian internal agricultural of production = 1802 lpcd using analysis period as 1997–2001. This study gives agricultural of consumption in Delhi = 1868 lpcd in the year 2011-12.

4. CONCLUSIONS

Food products (agricultural and livestock) have a major contribution (~ 90%) in WF of consumption for a person in Delhi. Rice and wheat products are the major contributors in crop water footprint due to high consumption. Our estimations showed that cereals' consumption could decrease by ~4.2% from 2011 to 2021. Overall consumption of livestock products is estimated to increase by ~24%, chicken consumption might increase by ~45%.

Results suggest, increase of 12.8% in total water footprint of Delhi from 2011 to 2021 is primarily because of direct and livestock water footprint increase. Direct water footprint is estimated to increase by 66% in this period. Results suggest indirect water footprint will have a marginal rise of ~7% from 2011 to 2021. It was found there is no significant water footprint contribution (less than 8%) of products essential in urban living as discussed in this report (in per capita scale). This is because the relative contribution of food products is much higher than the products required in urban living. 5 major food products were identified which contributed 73% of the total indirect WF and scenario based sensitivity analysis.

It was estimated 1243 MLD of virtual water was lost in Delhi by food wastage in the year 2011. It was estimated 11239 MLD of virtual water was lost in the year 2011 in production, storage and transportation phases. It was found virtual water lost due to inefficiencies in post-harvest practices such as storage and transportation is more than what Delhi needs (direct water) in a day.

It can be concluded enormous potential of water savings is there in Delhi as well as in India by reducing the losses in sowing, storage and transportation phases as well as by reducing the food wasted. It can ultimately lower the pollution levels in the streams by higher dilution.

For rural regions, emphasis should be on sowing seasonal crops according to climatic conditions, soil productivity and water availability. Cities can play a crucial role in decreasing indirect losses by connecting them with rural areas through efficient transportation systems. Proper storage facilities for agricultural products can be used as an indicator for sustainable cities and can aid in water savings.

REFERENCES

1. Ahams, I. C., Paterson, W., Garcia, S., Rushforth, R., Ruddell, B. L., & Mejia, A. (2017). Water Footprint of 65 Mid- to Large-Sized U.S. Cities and Their Metropolitan Areas. Journal of the American Water Resources Association, 53(5), 1147–1163. https://doi.org/10.1111/1752-1688.12563

2. Allan, J. A. (1996). Water Use and Development in Arid Regions: Environment, Economic Development and Water Resource Politics and Policy. Review of European Community and International Environmental Law, 5(2), 107–115. https://doi.org/10.1111/j.1467-9388.1996.tb00258.x

3. Bhagat, R. B. (2011). of Urbanisation in India. *Challenges*, *xlvI*(34), 10–12.

4. Central Ground Water Board. (2016). *Ground Water Scenario in India*. 1–39.

5. Chapagain, a K., & Hoekstra, a Y. (2004). Water footprint of nations. Volume 1 : Main report. *Value of Water Research Report Series*, *1*(16), 1–80. Retrieved from http://waterfootprint.org/media/downloads/Report16Vol1.pdf

6. Chapagain, A. K., Hoekstra, A. Y., & Savenije, H. H. G. (2006). Water saving through international trade of agricultural products. *Hydrology and Earth System Sciences*, *10*(3), 455–468. https://doi.org/10.5194/hess-10-455-2006

7. Feng, K., Hubacek, K., Minx, J., Siu, Y. L., Chapagain, A., Yu, Y., … Barrett, J. (2010). Spatially Explicit Analysis of Water Footprints in the UK. *Water*, *3*(1), 47–63. https://doi.org/10.3390/w3010047

8. Gleick, P. H. (2000). A look at twenty-first century water resources development. *Water International*, *25*(1), 127–138. https://doi.org/10.1080/02508060008686804

9. Gleick, P. H. (2006). Water and Conflict: Fresh Water Resources and International Security. *International Security*, *18*(1), 79. https://doi.org/10.2307/2539033

10. Gupta, S. (2013). *Ground Water Information - Greater Delhi*. 1–20.

11. Hoekstra, A. Y., Chapagain, A. K., & Aldaya, M. M. (2011). *The Water Footprint Assessment Manual*.

12. K Bharat, G., & B Dkhar, N. (2018). *DISCUSSION PAPER ALIGNING INDIA ' S WATER RESOURCE POLICIES WITH THE SDGs*. (November)

13. Kampman, D. (2007). The water footprint of India. *Value of Water Research Report Series*, (32), 152. Retrieved from http://www.utwente.nl/ctw/wem/education/afstuderen/afstudeerverslagen/2007/kampman.pdf

14. Mekonnen, M. M., & Hoekstra, A. Y. (2011). The green, blue and grey water footprint of crops and derived crop products. *Hydrology and Earth System Sciences*, *15*(5), 1577–1600. https://doi.org/10.5194/hess-15-1577-2011

15. NSSO. (2014). *Household Consumption of Various Goods and Services in India 2011-12*. *558*(558).

16. Rockstrom, J., Lannerstad, M., & Falkenmark, M. (2007). Assessing the water challenge of a new green revolution in developing countries. *Proceedings of the National Academy of Sciences*, *104*(15), 6253–6260. https://doi.org/10.1073/pnas.0605739104

17. United Nations. (2018). *World Urbanization Prospects: The 2018 Revision, Key Facts*. 2. https://doi.org/(ST/ESA/SER.A/366)

18. Vanham, D., Hoekstra, A. Y., & Bidoglio, G. (2013). Potential water saving through changes in European diets. *Environment International*, *61*, 45–56. https://doi.org/10.1016/j.envint.2013.09.011

19. Zhang, Z., Yang, H., & Shi, M. (2011). Analyses of water footprint of Beijing in an interregional input-output framework. *Ecological Economics*, *70*(12), 2494–2502. https://doi.org/10.1016/j.ecolecon.2011.08.011

20. URL1: https://waterfootprint.org/en/

21. URL2: https://niti.gov.in/content/voluntary-national-review-report

22. URL3: http://www.fao.org/faostat/en/#data/FBS

BIOGRAPHIES



Vaibhav,
Master of Technology Scholar,
Department of Environmental Engineering, Delhi Technological University



Dr Bharat Jhamnani,
Assistant Professor,
Department of Civil Engineering,
Delhi Technological University

Wireless Sensor Network Based Controlling and Monitoring for Smart Homes Using Zig Bee

Bhavnesh Jaint, S. Indu, Neeta Pandey

Abstract: ZigBee is a wireless communication protocol which finds application in home-automation environment monitoring, smart metering and industrial automation etc. Here wireless sensor network and ZigBee are utilized for smart home-automation. The proposed system monitors the electrical parameters e.g. voltage, current and power consumption. The novelty of this system is the implementation of a mechanism to control the appliances using a sensor network on ZigBee. The developed system is a low-cost, compact, and flexible in operation. Further it will help in reducing carbon footprint as it will allow the appliances to work only when desired and thus can save electricity expenses of the consumers. The prototype has been extensively tested in real-life situations and experimental results are very promising.

Index Terms: Home Automation, Monitoring, Wireless Sensor Network, ZigBee.

I. INTRODUCTION

The leading applications of wireless sensor network (wsn) are house monitoring and automation where a number of mixed sensors are installed to find out different actions of users. In WSN, many nodes which include sensors, actuators, controllers, and Radio frequency chips for wireless communication are used. First international standard for WSN IEEE 802.15.4 standard implemented in 2003 [1]. ZigBee communication protocol IEEE 802.15 was standardized by ZigBee Alliance and its Mesh network along with IPv6. Encryption, verification of network nodes, an effective and modern routing that leads to mesh. The WWW services is used from a smart house view point to connect households, objects, mainly electrical and electronic appliances with sensors and actuators by internet of things (IOT). The interests to create and outline the smart house are; enhancing extravagance, openly living, security and well-being, and successful use of power. To discover the issues and difficulties of the real execution of efficient home vigilation, we need to know present and progressing research here. A scope of brilliant home frameworks for encompassing helped living has been anticipated and grown, however there are, similarly few houses that utilize savvy advancements. Aware Home Research Initiative (AHRI) at GIT (Georgia Institute of Technology) [2], CASAS (The Center for Advanced Studies in Adaptive Systems) at Washington State University [3], AgingMo at University of Missouri [4], PlaceLab at MIT [5] and Smart home Lab at Iowa State University [6],[7] are checking the activity of everyday living, enhancing the solace by creating mindful relative circumstances through heterogeneous sensor positioned appropriately. These sensor units are accompanied with cameras as well. Electronic House by McDonough in New York, Toyota dream house in Japan, W.S at Stuttgart, and Crystal House by Hung at Taichung, Taiwan, House R128 by Intgeniere, in Germany which have focused on vitality protection objectives and maintainability and solace and offer accommodation through the mental aptitude of nature [8]. Many activities including the use of flag helped wearable gadgets are proceeding in various research labs Senswear Armband by Body Media Inc [9], SmartShirt [10], and PROETEX extend by CSEM focus [11], LifeShirt by Vivometrics [12] are wearable gear to quantify ECG. Other research has focused on fall location utilizing wearable sensors [13] in view of accelerometers. In this paper we developed a system which is low-cost, compact, and flexible in operation, also help in reduce in energy consumption thereby reducing carbon footprint.

The remaining sections of this paper are organized as follows. Section II related to literature review on automation in home appliances. In Section III discusses system architecture. Section IV related to working of system. Section V explain the controlling of electrical appliances. Section VI describe the hardware and software implementation. Section VII discusses the result of smart home automation section VIII conclude the paper.

II. LITERATURE REVIEW

The system explained in [14]- [16] provides different ways to control home appliance such as the Global System for Mobile (GSM), Zigbee, Bluetooth and internet. The microcontroller PIC16F887 is used to control the home appliances with GSM communication modulator and demodulator receives instruction from the mobile phone via GSM network [17]. It is a Short Message Service (SMS) based system. In system [18] is a Machine to Machine system and communication occurs through the GSM module. GSM has option for Machine to Machine communication which include the Dual Tone Multi-Frequency (DTMF), SMS and General Packet Radio Service (GPRS). The server at the home is based on the SMS/ GPRS mobile cell module and controller [19]. This can be used by the user to monitor and control the different home appliance at the home by using the Java enabled cell phone.

Revised Manuscript Received on March 25, 2019.

Bhavnesh Jaint, Department of Electrical Engineering, Delhi Technological University, Delhi, India.

S. Indu, Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, India.

Neeta Pandey, Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, India.

A voice operated control system is proposed [20] that used to enable or disable the home appliance remotely. The complete system utilized [21] the GSM communication network and all peripheral controlled by the Microcontroller of AVR. The proposed system was based on the SMS. The client provides the commands sent via SMS. The application of the Bluetooth technology and cell phone is to manage and control the home application [22]. The Personnel computer that is interface via Universal Serial Bus to the Bluetooth card, sensing network and a PWM circuit. The main disadvantage is that it has very limited range of operation about 10 meters only. Power point control using wireless technology hardware detailed given [23]. The main objective of this proposed network to manage the power of all devices and instruments at remote places. The instructions are sent through the telephone lines. The DTMF technique is provided by the telephone's lines [24]. The ZigBee Home application appliances can be used wirelessly [25]. Voice recognition and controller system is used for this aim. Wi-Fi communication technology [26]and Arduino board is used in this system. Light and temperature sensor can be used to control the home appliances. A detailed explanation of different technologies that can be applied to home appliance automation is done [27]. It provides the basic idea how modern technology networking can be applied for the home appliance applications. [28]. This paper presented a ZigBee based wireless air-conditioner remote controller architecture. [29] Design of an embedded system which improves the energy utilization rate, reduces the waste of energy consumption. In [30] the proposed study focuses on the usages of ZigBee and WSN in smart home automation system. In this, a Simulink model of different levels of energy consumption in household electrical appliances for monitoring consumption rate numerically and graphically is designed.

Table I Consolidated comparison of all systems:

System	Primary Communication	Remote Access	Number of devices	Speed	Real Time
GSM Communication	SMS	world wide access	Unlimited	Slow due todelivery issue	No
Bluetooth	Bluetooth and Attention commands	Restricted to range 10m	Unlimited	Fast due to proximity	Yes
Phone based	Phone lines	world wide with phone line	12 due Freq of DTMF	Fast	No
Zigbee	Zigbee and Attention commands	Around 10m	Unlimited	Fast	Yes
Wireless	IR or Other waves	Depending on range and spectrum of waves used	Unlimited	Slow due to interference	Yes

III. System Architecture

The system has been designed for the Control of connected household appliances and measurement of their electrical parameters remotely. Salient features of the system are the easiness of modeling, setup and user friendly. Electrical power consumption of various appliances in a house along with drawn current and supply voltage is the key parameter. Fig. 1 demonstrates the operational explanation of the proposed designed and developed system to control apparatuses in view of the consumer prerequisites and monitor electrical parameters. The electrical parameters of home appliances have been measured by interfacing with fabricated sensing modules. The functional details of the design and development of the sensing modules are explained in the following sections.

- Master Unit
- Slave unit

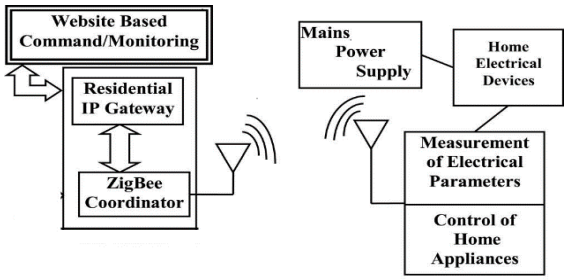


Fig.1 Functional Block Diagram of the System

The output signals from the sensors are integrated and connected to Microcontroller and ZigBee module for transmitting electrical parameters data wirelessly.

Master Unit: The master unit uses ATMega 328 Microcontroller based Aurdino as main controlling device which is interfaced with the following modules - Wi Fi, Zigbee, SD card, RTC and LCD. Block diagram of Master unit is shown in Fig 2. The level shifter in block diagram is used to provide appropriate power supply for the devices used as the power supply requirement are different for Wi Fi, ZigBee and Microcontroller.

The coordinator ZigBee Module communicates with Master Microcontroller and Slave ZigBee module. The SD card module stores the Data like voltage, current and power of each connected device with slave unit. The Real Time Clock (RTC) is used to generate time stamping for each log so that the exact time instance is known. The LCD displays present status of device (ON/OFF) with (1/0), voltage, current and power. LM2596 & LM1117 voltage regulator are used to generate fixed 5V and 3.3V supply. The microcontroller, **Level shifter**, RTC, SD and LCD module operate on 5 Volt whereas 3.3 volt is used to operate ZigBee, **Level Shifter** and Wi-Fi module.

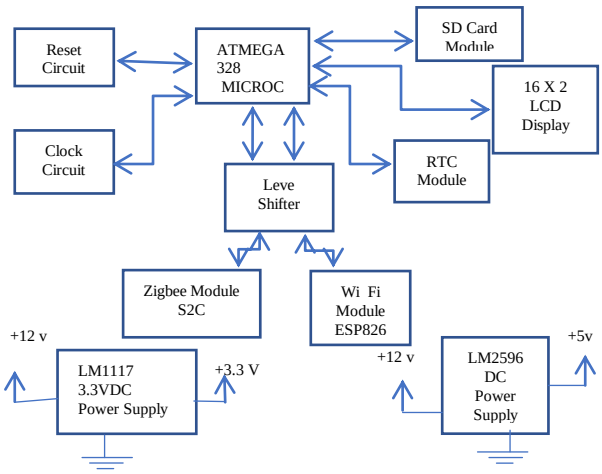


Fig.2 Block Diagram of Master Unit

Slave Unit: The microcontroller communicates with variety of sensors such as voltage sensors, current sensors and temperature sensors, along with switching circuit, ZigBee module and Power cutoff circuit. The voltage sensor includes a step down transformer that converts the voltage from 220V to 12V which is passed through voltage divider circuit and further converted to a digital signal by the ADC which is displayed on the LCD of Master Unit. Current sensors (ACS712) are placed in series of Loads. Microcontroller processes its values and display it on LCD of Master Unit. Temperature sensor (LM35) is used to measure the temperature of Slave unit area. In switching circuit, Solid State Relay (SSR) is used to switch ON/OFF the devices. Microcontroller sends a signal to SSR for switching purpose. It consists of Opto-coupler and the Triac. The Zigbee Module communicates with Slave Microcontroller and Coordinator Zigbee module and receives the status of each device from the slave microcontroller. It also receives the On/OFF instructions from the coordinator ZigBee. Its send the received slave microcontroller data to coordinator ZigBee and Coordinator ZigBee data to slave Microcontroller. LM7805 & LM1117 voltage regulator are used to generate fixed 5V & 3.3V supply respectively.

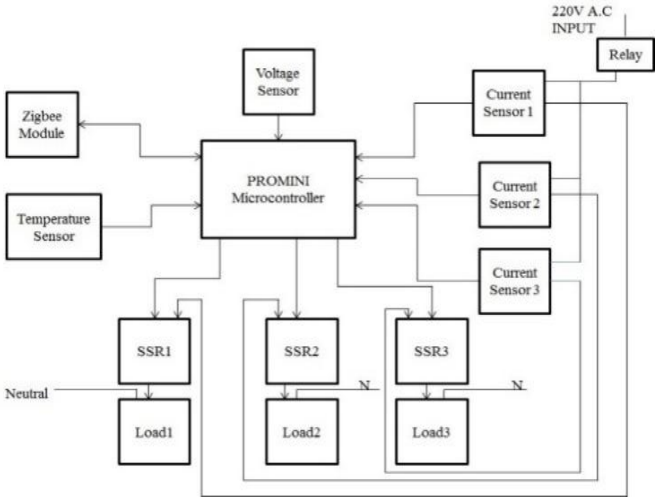


Fig. 3 Block Diagram of Slave Unit

IV. WORKING OF THE SYSTEM

The three Web pages, User / Admin login page, User web page and Admin web page were developed for complete operation of the system. Login user ID and Password are required to access User web page from the web login page. In user web page first check the status of the connected devices then user press the ON switch of device 1, which results into a command through MQTT to Wi Fi module, if Wi Fi module is connected to authorize internet connection then Wi Fi forward this command to Microcontroller, which sends a signal to Coordinator Zigbee module (Master). The coordinator Zigbee further sends command to Slave Zigbee, Slave Zigbee send a command to Slave microcontroller, the microcontroller generates a trigger command to SSR and which activates device 1. Voltage and current sensor interfaced with slave microcontroller, Slave microcontroller senses the voltage and current and sends to slave Zigbee, slave Zigbee sends status of devices, voltage and current to the master controller, which forwards all these signals to LCD. LCD will display status of devices, voltage, current and power of each device. The sensed data is also displayed on the user web page for monitoring and controlling the devices. A SD card is used to store information that comes from Master microcontroller. The stored data is used for data analysis for Load Management and other uses. A scheme has been setup where the system figures out the peak hour of electricity usage and it accordingly controls power consumption at home by switching off the un-important appliances. The system is connected to the mains (220- 240 V, 50 Hz) through monitoring circuit.

V. CONTROL OF ELECTRICAL APPLINCES

Our system is unique from other similar literature because of its feature of controlling. To give flexibility to consumer in controlling the device triac-BT139 along with intelligent Metering system has been used for switching device on/off. The user can turn the device off and on three different ways.

A. Automatic control

Our system can be programmed in such a way that it can automatically scrap webpage of electricity distributor of the particular user to get the tariff rate of electricity. And the system has the capability of deciding the peak hour and accordingly it will switch off or switch on the connected device, giving user some sort of automation

B. Manually control

The user is provided with a manual control of system. A OFF/ON switch is directly provided to user to control the system at his/her own wish. Our system is designed in such a way that a manual command will always override the automatic command

C. Remotely control

Users can remotely control the system from a remote location. It can be controlled from a webpage or even from an app. This is pretty useful for users who often forget to switch off their electric appliances before stepping out of home. The webpage will show the user about the current status of appliance and the user can turn the appliance on or off from the webpage

VI. HARDWARE DESIGN

The complete system has been designed and developed hardware wise in two units, called Master Unit and Slave Unit. Following Hardware Components Used in these Units.

- AVR ATMega328 Microcontroller working on Arduino UNO Platform
- Wi Fi Module - ESP 8266
- ZigBee Module 2C
- DC-DC Converter - LM 2596
- Voltage Regulator - LM 7805 & LM 1117
- Triac with optocoupler - BT139
- Current Sensor - ACS712
- Micro SD card Module
- RTC - DS 1307 module
- Level Shifter
- 16 x 2 LCD MATRIX
- Temperature sensor - LM 35

A. Hardware setup for master and salve unit

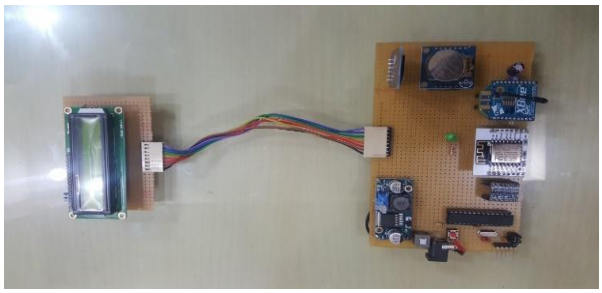


Fig. 4 Setup of Master Unit

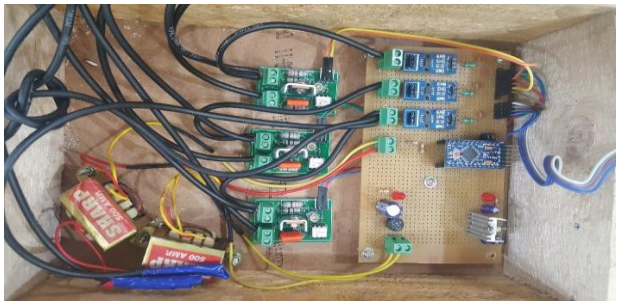


Fig.5 Setup of Slave Unit

B. Software implementation

Development of Web pages

For remotely control and monitor of electrical appliances, the three web pages have been developed using HTML, Java script, PHP. Login Web page has been developed for secure login in the internet for user and Admin. For login, user name and password are required. The Fig.6 has been given below for login the user or Admin.



Fig. 6 Login Page of user

In Fig.6 user name and password is entered. Here two pages was developed for Appliance user and Admin. Appliance user can control and monitor the all appliances which is connected through the developed Slave system. But admin can only monitor and ON or OFF the all appliances at a time in emergency condition.

After proper authentication by verifying the user name and password, the page with the information about remotely operating the three electrical appliances can be accessed. We also measure the voltage, current and power of each appliance and Temperature of Slave unit. Every status symbol represents its corresponding appliance condition. Here status symbol colour is blue, which indicate Master unit is not connected with internet hence cannot operate or monitor the appliances. When Master unit is connected with internet the colour of status symbol will change Red or Green. Red indicates particular appliance is OFF. Green indicates particular appliance is ON condition.

In the Fig.7 Admin page has been shown. This page gives information on the input voltage, current, Power of all appliances along with this the Temperature of Slave unit is also monitored.



Fig.7 Admin Webpage

Here also check the status of each device by the colour. Blue colour means Master unit is not connected with the internet. Red colour shows that appliance in OFF condition. While Green colour of status shows device in ON condition. Here Admin can OFF or ON all the devices in the emergency or as per requirement.

VII. RESULTS

The system designed here is best suited for smart homes. Our system tested on loads like Air Conditioner, Geyser, Television, Refrigerator, Water kettle, Toasters, Microwave oven etc. However, any number of electrical appliances can be put as load provided each load is below 4000W. AC operation and Monitoring of electrical parameters was tested using developed system has been shown in Fig.8. The AC was connected to Device 2 position. It was operated remotely through internet (webpage), the status of Device 2 will change from Red to Green. The web page also shows the input voltage 228 Volt, Current drawn by AC 6.284 Ampere and power consumption 1435 Watt. This also shows the temperature of slave unit 30°C.

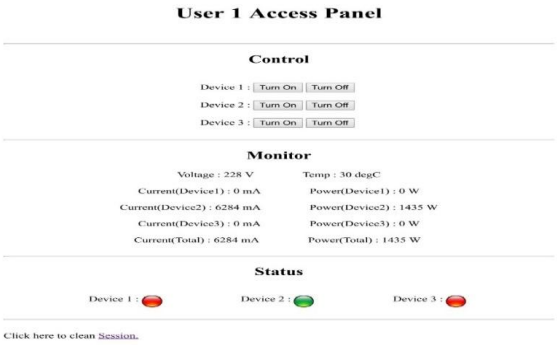


Fig.8 Controlling and Monitoring of Air Conditioner

Electric Iron operation and Monitoring of electrical parameters was tested using developed system has been shown in Fig.9.

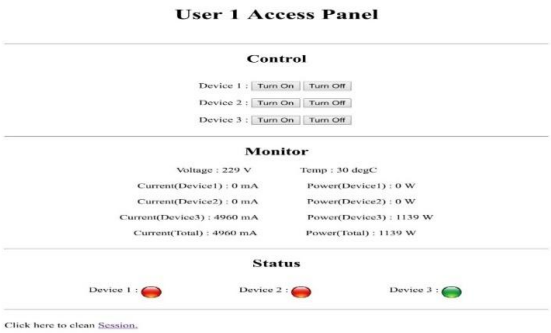


Fig.9 Controlling and Monitoring of Electric Iron

The Electric Iron was connected to device 3 position. The Iron was operated remotely through internet (webpage). When Iron operated, the status of Device 3 will change from Red to Green. The web page also shows the input voltage 229 Volt, Current drawn by AC 4.96 Ampere and power consumption 1139 Watt. This also shows the temperature of slave unit.

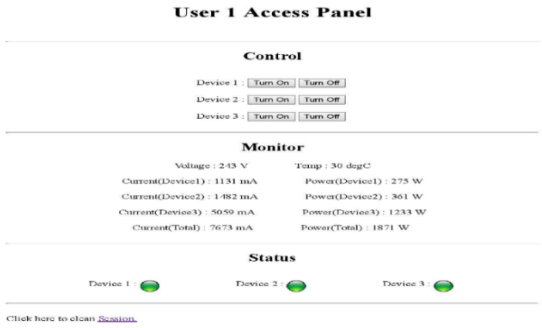


Fig.10 Controlling and Monitoring of all three devices

All three Devices were connected with the developed system for controlling and monitoring of electrical parameters of Water cooler, Freeze and Electric Iron shown in Fig.10. All the three devices connected respectively at Device 1, Device 2 and Device 3. All the devices were operated remotely through internet (webpage).

Table 2: Device status and current or power consumption

Device	status	current	Power consumption
1(water cooler)	on	1.131Amp.	275W
2(freeze)	on	1.482Amp.	361W
3(iron)	on	5.06Amp.	1233W
Total	All device on status	7.673 Amp	1871W

The system warns regarding failures to users. We have designed the system in such a way that it would automatically sense the peak hour of power consumption and it would automatically turn off the devices which are not in use at that particular time and a notification is sent to the user.

The system has been tested on many household electrical appliances and the results that were achieved are shown in the result sections. The 0 to 5 % error for all measured parameters. It can be seen that maximum error for household appliance is less than 5%.

VIII. CONCLUSION AND FUTURE SCOPE

An intelligent system that faithfully monitors and controls electrical appliance, for smart homes and intelligent buildings has been developed. A continuous watch on electrical appliances can be kept through a website. By an extension of these ideas, we can keep a watch on the entire building or City. Our aim here is to come up with an optimized low-cost solution that minimizes electricity consumption during peak hours. In future, this framework will be part of user's day to day life. People will be aware of possible power wastage and help reduce carbon footprint.

REFERENCES

1. IEEE 802.15™: Wireless Personal Area Networks (PANs). Available online:(accessedon1January 2015) http://standers.ieee.org/about/get/802/802.15.html
2. C. Bell; C. Fausset; S. Farmer; J. Nguyen; L.Harley; W.B. Fain; "Examining social media use among older adults*". In Proceedings of the 24th ACM Conference on Hypertext and Social Media*, Paris, France, 1–3 May 2013; pp. 158–163.
3. P. N. Dawadi; D.J Cook; M.S.Edgecombe, M.; C. Parsey, "Automated assessment of cognitive health using smart home technologies. *J. Tech. Health Care*, 2013, 21,4, 323–343.
4. M.Rantz, M.; Alexander, G.; Galambos, C.; Vogelsmeier, A.; Popejoy, L.; Flesner, M.; Lueckenotte, A.; Crecelius, C.; Zwygart-Stauffacher, M. Initiative to test a multidisciplinary model with advanced practice nurses to reduce avoidable hospitalizations among nursing facility residents. *J. Nurs. Care Qual*. 2013, 29, 1–8.
5. Place Lab MIT. Available online: http://web.media.mit.edu/~kll/A_PlaceLab_Sept1-2004.pdf (accessed on 10 January 2015).
6. M. Heinz; P. Martin; J. A. Margrett; M. Yearns; W. Franke; H. I. Yang; J. Wong; C. K. Chang; "Perception of technology among older adults" *J. Gerontol. Nurs.*, 2013, 39, 42-51.
7. H. I. Yang; R. Babbitt; J. Wong; C. K. Chang; "A framework for service morphing and heterogeneous service discovery in smart environments. In Impact Analysis of Solutions for Chronic Disease Prevention and Management", *International Conference on Smart Homes and Health Telematics*, Springer: Berlin/Heidelberg, Germany, 2012, 9–17.
8. S.Y Chen; S.F. Chang; "A review of Smart Living space development in a cloud computing network environment". *Comput. Aided Design Appl*. 2009, 6, 513–527.
9. Introducing the BodyMediaSenseWear System. Available online: http//www.sensewear.com/ solutions_bms.php (accessed on 12 January 2015).
10. D. Marculescu; R. Marculescu; N.H. Zamora; P. Stanley-Marbell; P.K. Khosla; S .Park; S. Jayaaraman; S.Jung; C. Lauterbach; W. Weber; "Electronic textiles: A platform for pervasive computing", IEEE Proc. 2003, 91, 12, 1995–2018.
11. G. Voirin; "Working Garment Integrating Sensor Applications Developed within the PROeTEX Project for Firefighters" *Ubiquitous Computing in the Workplace, Advances in Intelligent Systems and Computing*, 2015, 333, 25-33.
12. P. Grossman; "The LifeShirt: A multi-function ambulatory system monitoring health, disease, and medical intervention in the real world", *Studies in health Technology and informatics*. 2004, 108, 133-141.
13. P. Magaña-Espinoza; R. Aquino-Santos; N. Cárdenas-Benítez; J. Aguilar-Velasco; C. Buenrostro-Segura; A. Edwards-Block; A. Medina-Cass; "WiSPH: A Wireless Sensor Network-Based Home Care Monitoring System"; *Sensors;* 2014, 14, 7096–7119.
14. A. B. Yuksekkaya; M. A. Kayalar; M. B. Tosun; M. K. Ozcan, and Ali Ziya Alkar; "A GSM, Internet and Speech Controlled Wireless Interactive Home Automation System"; *IEEE* Transactions on Consumer Electronics, 2006 , 52, 3, 837 - 843.
15. K. Y. Lee ; J. W. Choi; "Remote Controlled Home Automation System via Bluetooth Home Network", , *SICE Annual Conference*, Fukui, 200, 3, 2824 – 2829.
16. F. Baig; S. Beg; M.F. Khan; "ZigBee Based Home Appliances Controlling Through Spoken Commands Using Handheld Devices", *International Journal of Smart Home*, 2013, 7,1, 19 -26.
17. R. Teymourzadeh; S. A. Ahmed; K. W. Chan and M. V. Hoong; "Smart GSM Based Home Automation System", *IEEE Conference on Systems, Process & Control*; Kuala Lumpur, Malaysia; 2013, 306-309.
18. A. Alheraish; "Design and Implementation of Home Automation System", *IEEE Transactions on Consumer Electronics*; 2004, Vol. 50 issue 4,pp. 1087-1092.
19. M. V. D. Werff; X. Gui and W. L. Xu; "A Mobile based Home Automation System, Applications and Systems"; *2nd International Conference on Mobile Technology*, 2005, Guangzhou, pp.5.
20. F. Baig; S. Beg; M.F. Khan; "Controlling Home Appliance Remotely through Voice Command", *International Journal of Computer Applications*, 2012, Vol. 48(17), pp.1-5.
21. S. R. Bharanialankar; C. S. ManikandaBabu; "Intelligent Home Appliance Status Intimation Control and System Using GSM", *International Journal of Advanced Research in Computer Science and Software Engineering*, 2014, Vol. 4, issue 4, pp.554 – 556.
22. R. Piyare; M. Tazil; " Bluetooth Based Home Automation System Using Cell Phone", *IEEE 15th International Symposium on Consumer Electronics*, 2011, Singapore, pp. 192 – 195.
23. H. B. Stauffer; "Smart Enabling System for Home automation", *IEEE Transactions on Consumer Electronics*, 1991, Vol. 37 (2), pp. 29-35.
24. E. M C Wong; "A Phone Based Remote Controller for Home and Office Automation", *IEEE Transactions on Consumer Electronics*, 1994, Vol. 40(1), pp. 28-34.
25. V. S. Narayanan; S. Gayathri; "Design of Wireless Home Automation and security system using PIC Microcontroller", International *Journal of Computer Applications in Engineering Sciences*, 2013, Vol. 3 (Special Issue), pp. 135- 140.
26. B. M. O. Al-thobaiti; I. I. M. Abosolaiman; M. H. M. Alzaharani; S. H. Almalki; M. S. Soliman; "Design and Implementation of a Reliable Wireless Real-Time Home Automation System Based on Arduino Uno Single-Board Microcontroller", *International Journal of Control, Automation and Systems*, 2014, Vol. 3(3), pp. 11 – 15.
27. A. R. Delgado; R. Picking; V. Grout; "Remote-Controlled Home Automation Systems with Different Network Technologies", *Centre for Applied Internet Research*, April, 2015 University of Wales, UK.
28. L. Zhang; R. Yan; H. Sun and Y. Qian; "A wireless General Air Conditioner Remote Controller for smart homes" *11th International Conference on Sensing Technology (ICST)*, 2017.
29. S. Wang; Y. Wang; M. Dai; "Development of monitoring system for thermal Energy consumption in intelligent Home"; *IEEE Chinese Automation Congress (CAC)*; China,. 30 Nov.-3 Dec., 2018.
30. E. Irmak; A. Kose; G. Gocmen; "Simulation and ZigBee based wireless monitoring of the amount of consumed energy at smart homes" *5th international Conference on Renewable Energy Research and Applications, Birmingham*, UK, 20-23 Nov. 2016.

AUTHORS PROFILE



Bhavnesh Jaint received her B.E. degree in Electronics and Telecommunication Engineering from Government College of Engineering, Jabalpur, M.P India and M.E. degree in Electrical Engineering from Jabalpur Engineering College, Rajiv Gandhi Proudhyogiki Vishwavidyalaya Bhopal, M.P., India. During her M.E., she has completed dissertation in the area of Fuzzy logic. She has started her carrier as Lecturer of Electronics and Communication at BBITRC, Jhangribad affiliated to UPTU. She has also worked as Assistant Professor at ABES Engineering College, Ghaziabad.

Presently, she is working as Assistance Professor in the Department of Electrical Engineering at Delhi Technological University, Delhi, India. She has taught different UG and PG Level courses in different institutions particularly at Department of Electrical Engineering. She has guided four M.Tech. students in the area of control system, Internet on Things. She is a member of different statutory committees of the university. She is also the life member of profession societies Indian Unit for Pattern Recognition and Artificial Intelligence. Also, a member of IEEE.

Her research interests include wireless sensor network, Embedded System IOT, control System.



Dr S. Indu received her B.Tech. degree in Electrical Engineering from TKM College of Engineering, University of Kerala, India, M. Tech. degree in Electrical Engineering from College of Engineering Trivandrum, University of Kerala, India and PhD in the area of Visual Sensor Networks from University of Delhi, Delhi, India. She Joined Electronics and Communication Engineering Department of Delhi College of Engineering in 1999 as a lecturer. Currently she is working as Professor and Head of the Department of ECE Department of Delhi Technological University. She has taught various courses of ECE Department at UG and PG Level. She has guided around 30 M.Tech. students. There are 6 Ph.D. students pursuing Ph.D. under her. Under her guidance 2 candidates completed PhD successfully.She is a Life Member of ISTE, IETE and IUPRAI. and senior member of IEEE.

Her area of research interest is Computer Vision, Sensor Networks and Image Processing. She has published one book "Smart Camera Networks" and one Book Chapter "Video Surveillance" published by INTECH, UK. She has published around 70 publications in international journals, National and International conferences. She has completed 3 Sponsored projects as Principal Investigator, sponsored by Naval Research Board and Department of Science and Technology.



Neeta Pandey received the M.E. degree in microelectronics from the Birla Institute of Technology and Sciences, Pilani, India, and the Ph.D. degree from Guru Gobind Singh Indraprastha University, New Delhi, India. She has served in Central Electronics Engineering Research Institute, Pilani; Indian Institute of Technology, New Delhi; Priyadarshini College of Computer Science, Noida; and Bharati Vidyapeeth's College of Engineering, New Delhi in various capacities. Currently she is working as Professor of Electronics and Communication Engineering Department, Delhi Technological University, New Delhi. She has taught various courses of ECE Department at UG and PG Level. She has guided around 24 M.Tech. students. There are 8 Ph.D. students pursuing Ph.D. under her. Under her guidance 8 candidates completed PhD successfully. She has authored more than 150 technical papers in reputed national and international conferences and journals. Her current research interests include analog and digital VLSI design. Professor Pandey is a Life Member of ISTE and senior member of IEEE, USA.

The current issue and full text archive of this journal is available on Emerald Insight at:
www.emeraldinsight.com/0972-9968.htm

Xenic insurance company: ethical dilemma of an employee

Naval Garg
Delhi Technological University, New Delhi, India

Received 21 August 2019
Revised 9 October 2019
Accepted 9 October 2019

Abstract

Purpose – This paper aims to depict the ethical dilemma of an employee in an insurance company who analyzed the group health policy of a major private telecommunication company. He noticed striking discrepancies and reported the findings to his superior.

Design/methodology/approach – Case study methodology is used for this study.

Findings – This paper reported the ethical dilemma faced by the employee.

Originality/value – This is an original work to the best of the author's knowledge.

Keywords Health insurance, Insurance, Ethical dilemma, Insurance analytics

Paper type Case study

Mohan Sharma was a happy young man. As a recent graduate of a prestigious insurance school, he joined claim department of Xenic Insurance Company Limited. Mohan was best among his counterparts in the college and has achieved his Masters in Insurance with flying colors. His area of interest was health insurance, and he aspired to use all his learning of insurance analytics and actuarial science in developing an exhaustive but affordable health insurance product to serve bottom of pyramid. One of his projects titled "Health insurance products in sachet" won the first prize in business plan competition organized by the Finance Ministry of India. He got special appreciation by the chief guest of the function for his societal concern and insurance acumenship. His research article titled "Exploring the linkage between health insurance and poverty" also won accolades from industry experts. His academic brilliance and bright future prospects brought him job offers from leading private insurance players of the market. But he politely denied the offers and joined Xenic insurance, a public sector company, to pursue his dream of serving the downtrodden through his association with the government organization.

Established in 1940, Xenic Insurance Company in 2019 is the largest mutli-line general insurer of the India, providing a range of products for individuals, families, small businesses and mid-sized and large corporate houses and governments. The company offers both standard and customized products in nearly all areas of risk including motor, health, personal accident, shopkeeper items, household items, baggage, fidelity, fire, engineering, cash, crop, animal and events. It has gained Number 1 spot in the insurance market in terms of premium collection, market share, life insured and growth rate. The gross business of the company crossed Rs 23,000 crores in March 2018 and company aimed at crossing Rs 30,000

© Naval Garg. Published in *Rajagiri Management Journal*. Published by Emerald Publishing Limited. This article is published under the Creative Commons Attribution (CC BY 4.0) licence. Anyone may reproduce, distribute, translate and create derivative works of this article (for both commercial and non-commercial purposes), subject to full attribution to the original publication and authors. The full terms of this licence may be seen at http://creativecommons.org/licences/by/4.0/legalcode

Rajagiri Management Journal
Emerald Publishing Limited
0972-9968
DOI 10.1108/RAMJ-08-2019-0009

crores by March 2021. The company has been rated "A" (Excellent) by AM Best Company, a US-based most reliable credit rating agency. Company's strong financials as indicated by its solvency margin, reserves and net worth helped them to gain "AAA" (stable) rating from CRISIL. Xenic won "Golden peacock business excellence award" and "General insurance company of the year award" in 2018. It entered in the global market in 1999 and has expanded its business in more than 30 countries now. It employs more than 20,000 employees worldwide. The company attributes its success to a large pool of prudent underwriters, dedicated claim settlement teams and visionary fund investors.

Mohan Sharma was on cloud nine with his first assignment. New job relocated him to Mumbai where he got company accommodation within one month of joining. He proudly asked his widow mother and younger sister to shift with him. His initial posting was in claim department with job design of analyzing and suggesting corrective measures for highly claim prone group health policies. Group health policies are taken by big companies for their employees. Since, big corporate houses with large number of employees negotiates with insurance companies thus premium charged for such policies are significantly small in comparison with other health policies. Health insurance policies provide facility of reimbursement of hospitalization expenses incurred due to illness/sickness, accident or disease. Xenic Insurance Company offers four different types of health insurance products. "Individual Mediclaim Policy" provides independent and separate coverage for each individual covered under the policy. "Xenic Happy Family Floater Policy" is designed to cover all family members with one sum insured. For example, in Individual Mediclaim policy each family member may be covered separately for (say) Rs 5 lakh, whereas coverage of Rs 5 lakh floats across all family members in case of Family Floater policy. "Group Medical Policy (Standard)" is offered to small business entities, trusts, educational institutes, co-operatives and registered associations. It provides extended insurance coverage (certain riders of individual mediclaim policy are relaxed) to a group of employees at discounted rates. "Group Mediclaim Policy (Customized)" offers tailor-made insurance coverage to a predefined group of employees. Premium is charged on the basis of extent of coverage requested by the client. *(Table AI shows comparison of coverage under four types of policy)*.

While analyzing one such group health policy of a major private telecommunication company, Mohan noticed striking discrepancies. He observed that most of the claims pertain to maternity. Quickly he started to analyze the demographic composition of the insured group. "Oh My God! Nearly 60 per cent of insured are unmarried" he exclaimed. Sensing something wrong, he started to go deeper into the claim files. Later on, he observed that the majority of cases (delivery) pertain to small hospitals with no computerized bills, which seems quite strange in the case of Mumbai. Moreover, claims were processed without ensuring birth certificates of the newborn children. He also observed that no new born baby has been added to the policy. He recalled the terms and condition of the policy which allows addition of newborn baby without any premium. Now, he realized that there is something fishy. All factual and circumstantial evidences indicate some bigger conspiracy. He reported his findings to his branch manager who asked him to leave that file with him and start working on some other case. He expected appreciation from his senior but manager cold response shocked him. He did not receive any satisfactory response from manager even after repeated reminders. Meanwhile, he started to analyze claim files of other group health policies. He observed similar types of unexplainable items in the files. In one case, one insured was hospitalized 30 times in the past two months. In another instance, he observed that the all the claims registered under a particular policy in one particular month were treated in same hospital.

He reported every finding to his branch manager but manager never responded back. Dissatisfied by apathy of branch manager, Mohan approached Johnson, divisional manager with a formal report of all his findings. Johnson gave a patient hearing to Mohan and replied:

> Look Mohan, you are working in the right direction and your findings can help our company in a big way. But being a senior, I suggest you to take more time to explore. It will help you to substantiate your findings with more data and statistical analysis.

Johnson's words provided new lease of life for dejected and frustrated Mohan. He decided not to leave any stone unturn to bring corruption to light. Next morning, as soon as he joined the office, branch manager handed over his transfer order to him. He was transferred to another nearby branch under the same division manager with immediate effect. While he was wrapping his belongings, he noticed signature of Johnson on his transfer letter. While he was about the leave his office, his branch manager called him and said:

> Please do not take any personal grudges with you. You should have understood my silence. Leave it now. Focus on your future and do not indulge yourself in unnecessary controversy in future. All the very best for your future

He left the place hurriedly to join his new office. Adding salt to injury, he was asked to join pension department in the new branch. New assignment shattered him badly as he found no enthusiasm and interest in repeated routine task of pension disbursement. He could not tolerate looting of public money. It gave him sleepless nights. Somehow, he gathered enough courage to report the incidence to regional head. While he was waiting for regional manager, he met Sridhar, one of the alumni of his insurance school. Sridhar told him that he is working there for the past 15 years and is well-versed with workings of the company. Mohan shared everything with him. Sridhar gave him a patient hearing and offered a smile in return. He said, "Look folk, everyone shares the pie and nobody is ignorant over here. If you hit someone's stomach, he will definitely hit you back". With these words, Sridhar left the place. While he was pondering over his future course of action, he got call from his mother who reminded him to pay sister's fee, electricity bill and installment of car on time.

Appendix. Teaching note

Background

*Case study: Xenic Insurance Company: Ethical Dilemma of an Employee*is designed to be taught in BBA/MBA/PGDM/BTech/BA course or other courses that covers business ethics, professional conduct, corporate governance, corporate ethics, morality at workplace or similar ethics-related issues. The case is about an employee named Mohan Sharma from one of the fastest growing insurance companies. The employee is facing an imminent moral dilemma wherein he is unsure about his future course of action. Mohan, a bright student and post graduate in insurance, joined Xenic Insurance Company Limited, which is a leading public sector insurance company. His ultimate aim of life is to serve the society through developing an exhaustive but affordable health insurance product to serve bottom of pyramid. Accordingly, he declines job offer from private companies and joined government company in claim department. While analyzing claim files of Group Mediclaim Policy of a leading telecommunication company, he observes certain discrepancies. When he reports the same to his branch manager and then subsequently to his divisional manager then he is transferred to some other branch. He is the sole bread winner of the family and his mother and unmarried sister rely on him fully. Case ends with a moral dilemma for Mohan.

Coverage	Individual mediclaim	Family floater	GMC (Standard)	GMC (Tailor-made)
Cover	Individual	Family	Group	Group
Hospitalization expense	Yes	Yes	Yes	Yes
Pre hospitalization expenses	Yes (30 days)	Yes(30 days)	Ye (30 days)	Yes(As per request)
Post Hospitalization expenses	Yes (60 days)	Yes(60 days)	Yes(60 days)	Yes(As per request)
Pre-existing diseases	No	No	No	Yes
Maternity	No	No	No	Yes
New born baby cover	Yes (from the date of insurance)	Yes(from the date of insurance)	Yes(from the date of insurance)	Yes (from the date of birth of child)
Cashless facility	Yes (Limited)	Yes(Limited)	Yes(Limited)	Yes(As per request)
Pre-acceptance health checkups	Yes (above 45 years)	Yes (above 50 years)	Yes (above 50 years)	No
Dental treatment	No	No	No	Yes(As per request)
Naturopathy	No	No	No	Yes(As per request)

Source: Based on Coverage provided by Major Indian insurers

Table AI.
Coverage comparison of four policies

Learning objectives

After analyzing the case, students would be able to:

- identify the moral dilemma faced by an individual in the organization and understand personal and organizational implications of various course of actions;
- understand various organizational and extra-organizational mechanisms and networks through which individual's value manifest;
- differentiate between different approaches to business ethics like teleological, deontological and emotive approaches;
- understand the fact that personal ethics, personal need of the hour and organizational circumstances and needs may contradict; and
- understand various provisions of whistle blower act and its larger implications.

Assignment questions

Q1. Who is Mohan Sharma? Discuss his values, motivation, future aspirations and family conditions.

Q2. What should be the future course of action for Mohan? Discuss implications of each alternative available with Mohan.

Q3. What are the legal provisions against corruption as per Indian Penal Code? Also discuss statutory provisions of whistle blower protection.

Teaching plan

Introduction

Ask students to get into the shoes of Mohan.

Q1. Who are you? (5-10 min)

The question bears paramount importance and should be repeated again and again to reinforce the importance of 'getting into the shoes of Mohan'. Diverse responses including name, student, learner, human etc. should be negated till students get into the role of Mohan. Nobody can assume the mental and psychological challenge of Mohan until and unless one becomes Mohan. So, correct answer of the question is 'Mohan Sharma

Q2. Who is Mohan Sharma? Discuss his values, motivation, future aspirations and family conditions. (10 min)

Although question seems very simple and easily answerable, but its answer has deep ramifications. Honest and deep evaluation will help learner in understanding the personality, perception and motivation of main character of the case. Following are the few responses to the question "Who is Mohan Sharma". Students should be allowed to brainstorm to reveal maximum personality traits of Mohan.

- Employee of Xenic Insurance Company
- Victim of corrupt system
- A bright learner with a passion for excellence
- A brave man with never to die spirit (remind he has already raised his voice to branch manager and divisional manager and has decided to meet regional manager)
- A researcher or have research intent (published paper as student of PG)
- Most importantly, he is a man with a heart, which is filled with love, affection and compassion for needy people

Second part of the question is quite interesting to answer as values, motivation, future aspirations and family conditions are all interlinked and intermingled. Two- three facts inscribed in case study hints towards Mohan's strong commitment and feeling for downtrodden people.These facts include learning of insurance analytics and actuarial science for product development, his business plan and research paper's theme, he declines lucrative private sector jobs to serve public sector company. His academic achievements and extra-curricular credentials highlight his talent, hardworking intent, competitive spirit, theoretical expertise of subject matter and zeal for excellence. His family conditions are also a bit challenging. He is the sole bread earner of the family. His sister depends upon his for her studies. Additionally, Indian wedding also puts additional burden on bride's side. Also, he would also need some cuisine against probable health issues of his old-aged mother. Looting of public money leaves Mohan sleepless, it reflects his inclination for transparency, accountability and good governance. He seems to be a honest and a man with impeccable integrity.

Q3. What should be the future course of action for Mohan? Discuss implications of each alternative available with Mohan. (30-40 min)

Alternative 1: Leave the office of Regional Manager silently and without discussing his hardship. And concentrate on constructive work and wait for opportunity to fulfil his dream to serve the needy people.

Alternative 2: Meet Regional Manager and discuss whole incident informally (without formal complaint of corruption).

Alternative 3: Meet Regional Manager and give written complaint with all proofs.

Alternative 4: Come back silently and submit complaint to Chief Vigilance Commissioner and seek anonymity and protection under Whistle Blower Protection Act, 2011.

Alternative 5: Write an anonymous letter to Chief Vigilance Officer and mark various authorities like Prime Minister Office, Department of financial services (Insurance Division), Ministry of finance, CMD of the company, CEO of Telecommunication company, IRDAI and honorable Supreme Court of India.

Alternative 6: Meet branch manager and divisional manager and get involved in corrupt practice. Ask for his share of loot.

Alternative 7: Leave the job and use his talent, expertise and skills in a private company.

Now, students should be encouraged to discuss personal, organizational and societal implications of Mohan's decision. In general rewards for bringing corruption to light are letter of appreciation or recognition by higher officials but repercussions are numerous and are often dangerous. Personal repercussions may include transfer to far-flanged areas like north-east India, organizational witch hunting in the form of false charge sheets, audit and vigilance, denial of leaves, training and promotion opportunities, ostracism at workplace, bad behavior and victimization. Non-action may lead to loss of confidence, self-belief, self-efficacy, value degradation, loss of trust in company and system. Mohan may become indifferent towards such activities and may loss courage to speak up again. More than personal loss, students should be motivated to devote some time in exploring organizational repercussion like loss of a talented, young and enthusiastic human resource, demoralization of other employees due to grapevine, financial loss to the company, enticement for other employees who are susceptible to get engaged in corrupt practices.

Q4. What are the legal provisions against corruption as per Indian Penal Code? Also discuss statutory provisions of whistle blower protection.

Students should be encouraged to go through relevant statutory provisions of the same.

Corresponding author
Naval Garg can be contacted at: naval_garg123@yahoo.co.in

For instructions on how to order reprints of this article, please visit our website:
www.emeraldgrouppublishing.com/licensing/reprints.htm
Or contact us for further details: **permissions@emeraldinsight.com**