SCHOLARLY PUBLICATIONS

A CURRENT AWARENESS BULLETIN
OF RESEARCH OUTPUT

@DTU

(79th Edition)

JULY 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 **Seventy ninth** 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 **July, 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 generic watermarking model for object relational databases, ***7.1.Vidhi Khanduja*** and Shampa Chakraverty, Computer Technology, DTU

2. A Hybrid-Based Verifiable Secret Sharing Scheme Using Chinese Remainder Theorem, ***3.Om Prakash Verma***[1], ***6.Nitin Jain***[2] and S. K. Pal, [1]ECE and [2]IT, DTU

3. A Study on Engineering Behaviour of Geogrid Reinforced Clay, ***6.Prachi Lingwal***, Shristi Negi and ***6.Ankur Mudgal***, DTU

4. Adaptive enhancement of underwater images using multi-objective PSO, ***7.Rajni Sethi***[1] and ***3.Indu Sreedevi***[2], [1]IT and [2]Electronics, DTU

5. Analysis of structural parameters on sensitivity of black phosphorus junctionless recessed channel MOSFET for biosensing application, Ajay Kumar, Neha Gupta, ***3.M. M. Tripathi*** and ***3.Rishu Chaujar***, ElectricalDTU

6. APPLICATION OF RETROSPECTIVE COST ADAPTIVE CONTROL TO THE IEEE 13-NODE TEST FEEDER, SYED ASEEM UL ISLAM, ELIZABETH L. RATNAM, ***8.1.ANKIT GOEL***, and DENNISS.BERNSTEIN, Mechanical, DTU

7. Assessment of bacoside production, total phenol content and antioxidant potential of elicited and non-elicited shoot cultures of Bacopa monnieri (L.) , ***6.Nupur Jauhari***, Rakesh Bharadwaj, Neelam Sharma and ***3.Navneeta Bharadvaja***, Biotechnology, DTU

8. Development of Polypropylene/Banana Peel (treated and untreated) composites with and without compatibilizer and their studies, ***6.Priyanka Luthra***, ***3.Ram Singh*** and Gurpreet Singh Kapur, Applied Chemistry, DTU

9. Ecosystem Services by Wetlands for Polluted Water Treatment, Padma Vasudevan, P.K. Sen, R.K. Srivastava, Mamata Tandon, Shriram Hegde, ***3.T.Vijaya Kumar*** and Philip Davies, Civil, DTU

10. Effect of direct LN2 single jet supply during turning of AISI D3 steel alloy using an optimization technique, ***6.Anurag Sharma***, ***3.R.C. Singh***, ***3.Ranganath M.Singari***, Mechanical, DTU

11. EFFECTING PARAMETERS OF BIOGAS PRODUCTION, Ashok Maan and ***6.Pradeep Kumar Meena***, Mechanical, DTU

12. Efficient MDCT Recursive Structure for VLSI Implementation, ***6.Pragati Dahiya*** and ***3.Priyanka Jain***, Electronics, DTU

13. . Flame retardancy of ceramic-based rigid polyurethane foam composites, ***6.Anuja Agrawal[1]***, ***3.Raminder Kaur1*** and ***3.Ravinderjit Singh Walia***[2], [1]Applied Chemistry & Polymer and [2]Mechanical, DTU

14. Fractional Order High Pass Filter Based on Operational Transresistance Amplifier with Three Fractional Capacitors of Different Order, ***7.1.Gagandeep KAUR***, Abdul Quaiyum ANSARI and Mohammad Shabi HASHMI, Electronics, DTU

15. Hybrid Control Approach using NLMS and PLMS Algorithms for Grid Connected SPV System, ***3.M. Rizwan*** and ***6.Priyanka Chaudhary***, Electrical, DTU

16. INVESTIGATION OF LOCAL OPENING-SLIDING RELATIONSHIP IN THE VICINITY OF DEFORMED BAR IN CONCRETE BY USING DIC TECHNIQUE, A. OKEIL, ***7.R. UPADHYAY***, K. MATSUMOTO AND K. NAGAI, Mechanical, DTU

17. Investigation on flammability of rigid polyurethane foammineral fillers composite, ***6.Anuja Agrawal***, ***3.Raminder Kaur*** and ***3.Ravinderjit Singh Walia***, Applied Chemistry and Mechanical, DTU

18. Investigations of tribodynamic characteristics of chrome steel pin against plain and textured surface cast iron discs in lubricated conditions, ***3.Roop Lal*** and ***3.R.C. Singh***, Mechanical, DTU

19. Isolation Improvement of MIMO Antenna using a Novel Flower Shaped Metamaterial Absorber at 5.5 GHz WiMAX band, ***6.Priyanka Garg*** and ***3.Priyanka Jain***, Electronics, DTU

20. Kernel intuitionistic fuzzy entropy clustering for MRI image segmentation, ***3.Dhirendra Kumar***, R. K. Agrawal and Hanuman Verma, Appl Mathematics, DTU

21. Lithium-ion Battery Safety Tests monitored by LabVIEW and Comsol software for electrical and mechanical characteristics, ***6.Snigdha Sharma***[1], ***3.M M Tripathi***[1] and ***3.Amrish K Panwar***[2], [1]Electrical and [2]Applied Physics, DTU

22. Negative axicon tip micro-cavity with a polymer incorporated optical fiber temperature sensor, DNYANDEO PAWAR, AJAY KUMAR, RAJESH KANAWADE, SAMIR MONDAL AND ***3.RAVINDRA K. SINHA***, Applied Physics, DTU

23. On Random Read Access in OCB, ***8.1.Ashwin Jha***, Cuauhtemoc Mancillas-L´opez, Mridul Nandi and Sourav Sen Gupta, Computer, DTU

24. On-chip learning in a conventional silicon MOSFET based Analog Hardware Neural Network, Nilabjo Dey, Janak Sharda, Utkarsh Saxena, Divya Kaushik, ***7.Utkarsh Singh*** and Debanjan Bhowmik, Electronics, DTU

25. Optimal Q-Learning Approach for Tuning the Cavity Filters, Even Sekhri, Mart Tamre and ***3.Rajiv Kapoor***, Electronics, DTU

26. Printing of Lycell fabric with Rubia Cordifolia and Acacia catechu using Guar gum Chitosan as Thickening Agent, Divya Ojha, Sangita Deodiya and ***3.Roli Purwar***, Applied Chemistry, DTU

27. Problems of Scaling and Corrosion in Geothermal Water: Indian Scenario, ***8.Shivam Sharma***, ***8.Sarthak Aggarwal***, ***8.Yash Kumar Baghel***, ***8.Ajit Kumar*** and ***3.J. P. Kesari,*** Mechanical & Automobile, DTU

28. Role of Industry Factors in Financing the Outward Foreign Direct Investment by Indian Multinational Enterprises, Vanita Tripathi and ***3.Sonal Thukral***, DSM, DTU

29. Seasonal behaviour of thermal stratification and trophic status in a sub-tropical Palustrine water body, ***8.Priyanka Singh***, ***8.Jyoti Bagrania*** and ***3.A. K. Haritash***, Environmental, DTU

30. Studies on Polyacrylate-Starch/Polyaniline Conducting Hydrogel, ***6.Reetu Prabhakar*** and ***3.Devendra Kumar***, Applied Chemsitry & Polymer, DTU

31. Thermodynamic analysis of combined cycle power plant with trans-critical cycle integrated with solar system for power generation for space heating and cooling, ***3.R.S. Mishra*** and ***7.Aakash Behl***, Mechanical & Production, DTU

32. THERMODYNAMIC PERFORMANCE ANALYSIS OF DEDICATED MECHANICALLY SUBCOOLED VAPOUR COMPRESSION REFRIGERATION SYSTEM, ***3.S. Agarwal***, ***3.A. Arora*** and ***3.B. B. Arora***, Mechanical, DTU

33. Unified Graph-Based Multicue Feature Fusion for Robust Visual Tracking, Gurjit Singh Walia, ***6.Himanshu Ahuja***, ***8.Ashish Kumar***, ***3.Nipun Bansal***, and Kapil Sharma, Computer, DTU

34. What's Most Broken? A Tool to Assist Data-Driven Iterative Improvement of an Intelligent Tutoring System, ***8.Mononito Goswami***, Shiven Mian and Jack Mostow, DTU

35. Workplace Spirituality and Job Satisfaction: Exploring Mediating Effect of Organization Citizenship Behaviour, ***3.Naval Garg***, B. K. Punia and Anuradha Jain, Management, 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*

Multimedia Tools and Applications
https://doi.org/10.1007/s11042-019-07932-3

A generic watermarking model for object relational databases

Vidhi Khanduja[1] · Shampa Chakraverty[2]

Received: 28 September 2018 / Revised: 15 June 2019 / Accepted: 24 June 2019
Published online: 04 July 2019
© Springer Science+Business Media, LLC, part of Springer Nature 2019

Abstract

Object-Relational Databases have been implemented largely and thus, there is urgent need to protect them from potential misuse. Digital watermarking is considered as technological protection measure that deters the attempts to plagiarize or distort databases. Prior work focused primarily on protection of Relational Databases by digital watermarking. However, Object-Relational Databases seem to have escaped the attention of researchers. In this research work, we propose the framework for a novel, generic watermarking technique for ORDBs with a view to incorporate Ownership and Integrity protection. The software developed on the basis of this model is called ***WORD*** – Watermarking Object Relational Databases. To the best of our knowledge, this is for the first time that a complete model for watermarking ORDBs has been developed. To deal with changing security aspects with time, strategy pattern is adopted while implementing the watermarking model *WORD*. The use of strategy pattern designs incorporates flexibility and scalability in our model. Robustness and integrity of proposed system has been proved theoretically as well as practically by various experiments.

Keywords Object-relational databases · Digital watermarking · Database protection · Fragile watermarking

1 Introduction

With the Internet becoming part and parcel of our lives, there is an increase in vulnerability to copyright and piracy threats as well. Technological methods that are designed to protect digital

✉ Vidhi Khanduja
vidhikhanduja9@gmail.com

Shampa Chakraverty
Apmahs.nsit@gmail.com

[1] Computer Engineering Department, SAL Institute of Technology and Engineering Research, Ahmedabad, Gujarat, India

[2] Computer Engineering Department, Netaji Subhas University of Technology, Delhi, India

content such as images, video and databases are called "Technological Protection Measures" (TPM). They include the technologies that are used to control access to copyrighted digital content or to prevent users from copying such protected content. Digital watermarking is one such TPM that has emerged as an effective means to protect shared and outsourced databases from infringement. Watermarking protects digital content by embedding some digital marks and thus considered as a viable and cost-effective TPM. For web databases, the watermarking algorithms designed introduce small changes in the data being protected without altering the database in *significant way*. It provides a means to establish the true identity of the owner and deters attempts to plagiarize or distort it [1]. Prior state of art shows several works for protecting Relational databases (RDBs). However, Object Relational databases (ORDBs) are recently being implemented successfully in various domains and hence have escaped the attention of researchers. As ORDB works on objects rather than simple values, thus, existing techniques cannot be applied to them. In this research, we propose a watermarking model that taps unique characteristics of ORDBs to resolve two important security issues: Ownership proof and Tamper detection. We briefly explain the concept of ORDBs in Section 2.

Unlike encryption and hashing techniques, watermarking does not attempt to hide the data, but instead infuses a kind of ownership proof in the data. Encryption and hashing provide protection to the content by making the information indecipherable by an attacker. On the contrary, digital watermarking works on the principle of modifying the content in a manner so that the usability of the data is retained fully. In fact an attacker or observer has no way to decipher that a watermark is actually present in the database. However, the watermark does remain within the content inseparably providing a proof of the ownership or a signature to detect tampering or to track the people who may have obtained the content legally and are illegally redistributing it [17]. The main contributions of this paper are summarized below:

(i) Simultaneous treatment of both ownership proof and tamper detection in ORDBs by sequentially applying two different watermarks for ownership proof and tamper protection, and by using the services of a trusted third party.
(ii) The exploitation of two unique features of ORDB to enhance the security capabilities of our watermarking scheme: Firstly, the unique owner-accessible Object IDentifier (OID) of each row object is utilized for concealing the watermark, thereby making the scheme primary key independent. Secondly, the watermark is concealed in object type attributes giving us more bandwidth to accommodate watermark bits profusely, thus increasing the degree of robustness.
(iii) Development of WORD, an extensible software system based on the proposed ORDB watermarking model.
(iv) We used the principles of design patterns to make the software flexible and scalable so that new types of attributes can be incorporated as a database evolves.

The rest of paper is organized as follows: In Section 1 we have discussed about digital watermarking and its need to secure ORDBs. The details of features and applications of ORDBs are given in Section 2. In Section 3, we present prior sate of art. In Section 4, we describe the architecture of the proposed ORDB watermarking model. Section 5 analyses the security features incorporated in the proposed model. Section 6 illustrates the implementation details. Section 7 describes the experiments performed and demonstrates the results obtained on robustness and fragility of experimental database. Finally, we conclude the work in Section 8.

2 Background of object-relational databases (ORDBS)

ORDBs are relatively new on the horizon and have gained a lot of popularity due to their advantages over conventional relational databases [32]. An ORDB models object oriented components such as objects, classes and inheritance relationships into a relational database schema [9]. It provides support for including multimedia objects such as audio, video, image and graphics in the form of attributes. The ORDBs differ significantly with that of RDBs because of the following factors:

1. Data types in ORDBs differ significantly from that of RDBs. ORDBs are aimed at application domain where complex data plays significant role.
2. Number of potential location varies significantly.
3. In RDBs, primary key is used to identify the records uniquely, while in ORDBs there is an object identifier (OID) that refers to the inherited data.

Due to the differences, the technique to watermark ORDBs differs from that of RDBs. Several real-world applications are proof of their growing popularity. Barrodale computing services Ltd. has used ORDBMS to provide solutions to various applications involving the management of complex data types [3]. Some of the applications that use ORDBs include weather modelling and prediction system software, Commercial shipping information system, Life sciences databases and Ocean observatories.

An ORDB consists of groups of object tables. An object table is a special kind of table that holds objects called row objects. It gives a relational view of attributes of those objects. The allowable data-types in row objects are dependent or independent complex data types and primitive data types. A complex data type may be an object type, or collection of data elements of similar and dissimilar objects [25]. Objects that are not referenced by other objects are termed independent objects. In the proposed work, we conceal a watermark in an independent object type attribute of ORDB. Figure 1 illustrates an example of object type and object. *Emp _ Detail* is created as an object type that contains details of an employee. *Emp _ Detail* comprises of various attributes that make up a structured data unit. These attributes can be of primitive data type or another object type or an array of similar elements. *Emp _ Detail* includes *Emp _ ID* and *Name* as non-numeric type attribute; we refer such data type as string. *D _ O _ J* attribute maintains date of joining of employee. It is an object type that comprises of date i.e. day, month and year. *Tot _ Exp* attribute stores total number of experience in years an employee is having. We refer this data type as float indicating it contains fractional values. *PrevEmpDetails* attribute is object type describing previous employment details of employee. It comprises of attributes of float and string types described in *P _ Detail* object type. Object represents instance of object type *Emp _ Detail*. It shows the value of all the attributes in *Emp _ Detail*.

3 Literature survey

We now analyze the state-of-art in the domain of watermarking for digital databases. Prior work focused primarily on Relational Databases (RDBs). Khanduja et al. [14, 18], Halder et al. [8] have presented the detailed literature survey in domain of digital watermarking of

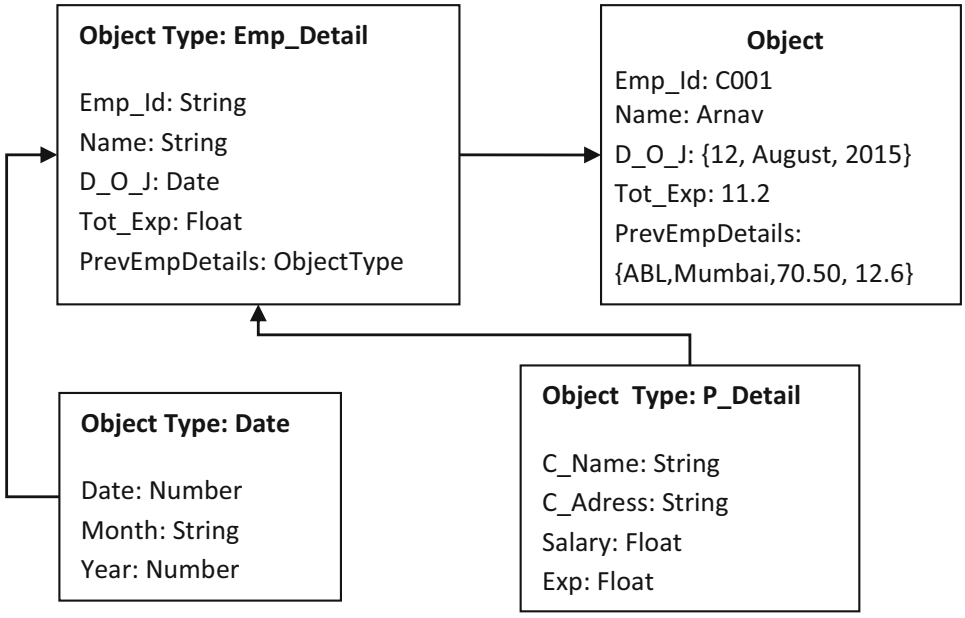


Fig. 1 Illustration of object type and object through an example

databases. We classify watermarking databases techniques into two categories: Robust watermarking techniques and fragile watermarking techniques.

i. **Robust Watermarking Techniques:** We apply robust techniques to resolve ownership issues. The watermark is embedded into a database redundantly so that watermark can be extracted successfully even after alterations introduced by an attacker. The main objective is to make a watermark robust against various attacks.
ii. **Fragile Watermarking Techniques:** Fragile watermarking techniques aim at tamper detection. The watermark usually acts as a signature of a database. Even the slightest alterations made to a database, deliberately or otherwise, will immediately alter its signature and hence its watermark. The tampering effect can therefore be detected. The main objective here is to make the watermark fragile to various attacks.

Table 1 represents the landmarks in state of art of robust watermarking techniques for the RDBs. The watermark can be created using random bit stream [2, 30], specific identifier [7], Image [33] or can be biometric trait such as voice [16]. The target attribute where watermark is inserted includes numeric [2, 7, 13, 16, 20, 30, 33, 35] or categorical [13, 20]. Date type column in Table 1 represents the target attribute data type.

Surprisingly, relatively fewer attempts have been made to develop schemes that protect the integrity of relational databases with fragile watermarking. They work upon the principle that the slightest change on data will immediately destroy the watermark, thereby detecting such attempts. Table 2 enumerates the work of researchers on protecting integrity of database. To the best of our knowledge no prior work has been reported on watermarking ORDBs which is the focus of the present research. In the proposed technique, the exploitation of two unique features of ORDB is done to enhance the security capabilities of our watermarking scheme: Firstly, the unique owner-accessible Object IDentifier (OID) of each row object is utilized for concealing the watermark, thereby making the scheme primary key independent. Secondly, the watermark is concealed in object type attributes giving us more bandwidth to accommodate watermark bits profusely, thus increasing the degree of robustness.

Table 1 Summary of robust watermarking techniques

Proposed schemes	Watermark content	Data type	Features
Agrawal et al. [2]	Random	Numeric	1. Identified need of watermarking Relational databases
Shehab et al. [30]	Random	Numeric	1. Partitioning without marker tuples 2. Optimization based technique.
Farfoura et al. [7]	Owner's identity	Numeric	1. Reversible technique 2. Use of Time stamping protocol
Sun et al. [33]	Image	Numeric	Multiple images are embedded
Khanduja et al. [16]	Biometric trait	Numeric	Use of Owner's voice as watermark
Li et al. [20]	Signature	Categorical	1. Distortion free technique 2. Localization up to group level
Khan et al. [13]	Signature	Categorical	Distortion free technique
Khanduja et al. [19]	Prepared using pivot attributes	Numeric	Information recovery is achieved along with robustness
Naaz F. [24]	Owner's identity	Zero Watermarking scheme i.e. watermark not embedded	1. Introduced watermarking as a service (WaaS)
Tufail H [35]	Random	Numeric	Watermark Creation through Binary Bat Algorithm (BBA)

4 Architecture of WORD

We proposed a watermarking model that provides ownership protection and detects any tampering done maliciously or inadvertently to ORDB. A database $\mathcal{O}$ is built upon Object relational model. The ORDB $\mathcal{O}$ consists of a group of object tables comprising N_t row objects. We assume a total number of fixed columns in a row object as N_a. To maintain simplicity in text, we ignore methods within an object as they don't play significant role in the watermarking process. The owner enumerates a set of secret parameters: secret key $\mathcal{K}_s$, row object selection parameterγ, number of candidate attributes β, number of permissible Least Significant Bits (LSB) within an attribute to embed a watermark α and a number of partitions N_p. These parameters are secretly decided by the owner and are kept private.

Figure 2 depicts a generic architecture of*WORD*. It consists of five processes. Firstly, we prepare a watermark $\mathcal{W}_o$ for ownership protection and in second process we profusely conceal it within a database $\mathcal{O}$. Next, we prepare signature of a database $\mathcal{W}_s$ for tamper detection and in forth process we register it with trusted third party (TTP). Lastly, *Generate Decision* phase

Table 2 Summary of fragile watermarking techniques

Proposed schemes	Watermark type	Target attribute	Contribution
Tsai et al [34]	Image	No embedding	Distortion free technique based on public authentication
Bhattacharya et al. [5]	Using database	Categorical	Distortion-free watermarking technique
Iqbal et al. [11]	Using database	Numeric	Self-constructing watermarking technique
Khanduja et al. [15]	Using Database	Numeric	Use of Rough Set theory.

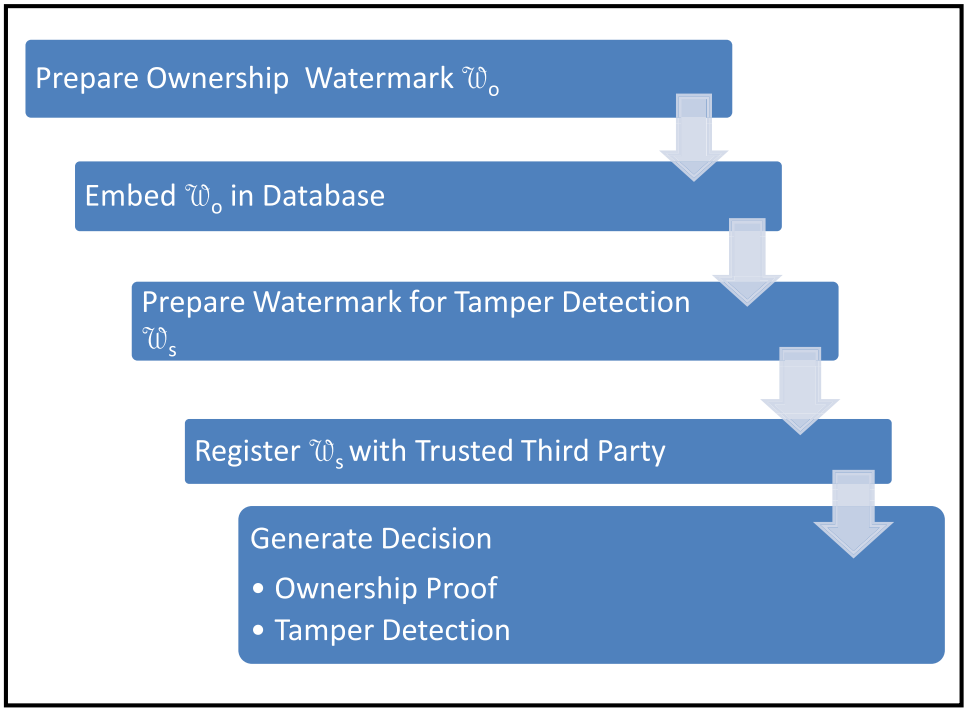


Fig. 2 The overall process of the proposed watermarking model *WORD*

decides true owner and/or whether database is tampered. Table 3 enlists the symbols used in this manuscript. We now discuss each of these steps in detail subsequently.

4.1 Prepare ownership watermark $\mathcal{W}_o$

This is the first process of *WORD* thatsecurely prepares a watermark as shown in Fig. 2. It converts the identity of a database's copyright holder to a watermark, thus ensuring proof of ownership. We prepare ownership watermark $\mathcal{W}_o$ using owner selected watermark w and secret key $\mathcal{K}_s$.

We follow the Kerckhoff's principle that says a system must be secure by using secret key $\mathcal{K}_s$; even if the algorithm used is publically known [12]. If the direct watermark i.e. say owner's identity is taken as a watermark; then an attacker can guess the watermark knowing the owner of a database. Use of secret key $\mathcal{K}_s$ makes this process secure.$\mathcal{W}_o$ is calculated using Eq. 1 as:

$$\mathcal{W}_o = hash\left(w \,\|\mathcal{K}_s\right) \tag{1}$$

Where, K_s is secret key, $\|$ is a concatenation operator, and hash(.) is a cryptographic hash function. A cryptographic hash function is a procedure that takes an arbitrary block of data and returns a fixed-size hash value such that any intentional change to the data will change its hash value [28]. Hash function provides the owner with a liberty to select his own watermark. Owner can select large files such as an image, audio or video as a watermark which are compressed to fixed size by applying a hash algorithm.

Many cryptographic hash algorithms exist in literature, e.g. RIPE-MD, MD5, SHA-2, SHA-3, HAVAL, SNEFRU, etc. We employ SHA-2 algorithm as a cryptographic hash function that yields 256-bit hash value owing to its improved resilience against attacks [26,

Table 3 Parameter table

S.No	Symbol	Representation
1.	$\mathcal{O}$	Object-Relational Database.
2.	N_t	Total number of Row Objects in database.
3.	N_a	Total number of attributes/columns in database.
4.	$\mathcal{K}_s$	Secret key decided by owner of database.
5.	γ	Row Object selection parameter i.e. number of row objects selected for watermarking within a database.
6.	β	Number of candidate attributes selected for watermarking within a database.
7.	α	Number of permissible Least Significant Bits within an attribute to embed a watermark.
8.	N_p	Number of partitions of the database.
9.	$\mathcal{W}_o$	Ownership Watermark for proof of ownership
10.	$\mathcal{W}_s$	Signature Watermark for Tamper Detection
11.	N_w	Total number of bits in Ownership watermark
12.	$OID(r)$	Object Identifier of row object r.
13.	$\mathcal{H}_o(r)$	Hash of row object r.
14.	$WIndex(r)$	Index of watermark bit to be inserted in r row object
15.	O_w	Watermarked Database
16.	$P = \{P_0, \ldots, P_{I_d}, \ldots, P_{N_p-1}\}$	A set consisting of $N_p$virtual partitions such that for any two partitions $P_i \cap P_j = \{\varnothing\}$, *if* $i \neq j$.

36]. Final watermark comprises of N_w binary bits that are securely embedded following a secure procedure explained in the next step.

4.2 Embed ownership watermark

This is the second process of *WORD*as shown in Fig. 2, where watermark $\mathcal{W}_o$ is embedded repeatedly into the object type attribute of the database.

An object type can be of primitive data type, multimedia data type or another object type or an array of similar elements. Thus, a single object type attribute has a potential to hide multiple instances of the watermarking bits within varied attributes present in it, thereby incorporating a higher degree of robustness. We now discuss, various sub-processes involved in the process of concealing $\mathcal{W}_o$.

A. **Select target positions:** Figure 3 shows the pseudo-code to calculate the potential locations for embedding $\mathcal{W}_o$. The pseudo-code commences by calculating row object's hash $\mathcal{H}_o(r)$ using Eq. 2.

$$\mathcal{H}_o(r) = Hash(OID(r)||\mathcal{K}_s) \tag{2}$$

Where, Hash(.) represents cryptographic hash function, $\mathcal{K}_s$ is secret key decided by the owner and *OID* is secure row object's unique identifier. Every row object is assigned a logical unique Object IDentifier (*OID*). No amount of modifications made to a row object will ever alter its *OID*. Moreover, *OID* is granted special privileges to make it visible only to the owner of the database. This feature makes it attractive to use *OID* in contrast to primary key used for

Select_Positions (.)

Input : $\mathcal{O}, \gamma, \beta, \alpha, \mathcal{K}_s, N_w$
Output : $\mathcal{H}_o(r)$, target positions $W_Status, AIndex, BPos, WIndex(r)$

1. **For** every row-object r ε $\mathcal{O}$
2. Calculate $\mathcal{H}_o(r)$using equation 3.2
3. Calculate $S_f = \mathcal{H}_o(r)[1:20]$, $S_e = \mathcal{H}_o(r)[1:30]$
4. **If** $\left(\left(S_f \ mod \ 1/\gamma\right) == 0\right)$
5. Set $W_Status(\ r) = 1$
6. Select embedding attribute: $AIndex = mod(S_e, \beta)$
7. Select $BPos = mod\left(\left(S_f || \mathcal{K}_s\right), \alpha\right)$
8. $WIndex(r) \ = mod(S_e, N_w)$
9. **End If** (line 4)**.**
10. **End for** (Line 1)

Fig. 3 Pseudo-code for selecting target positions to embed watermark

RDBs. The algorithm uses two different portions of the partition hash $\mathcal{H}_o(r)$ (line 3) for making the following decisions:

(i) $S_f(r)$– the first 20 bits $\mathcal{H}_o(r)[1:20]$, decides whether the row object r is selected for embedding a watermark bit or not (line 4). This process serves two purposes. One, it limits the amount of distortion by limiting the number of watermarked row objects. Secondly, the use of a secure hash function and secret parameters enhances security by concealing the identity of the watermarked row objects from an intruder.

(ii) $S_e(r)$– the first 30 bits $\mathcal{H}_o(r)[1:30]$, chooses the target attribute for embedding if r is selected for this purpose (line 6). Thus, for any given row object, one out of β candidates attributes is chosen randomly for embedding. This process adds another level of obfuscation by ensuring that the same attribute is not always chosen across different row objects for the purpose of embedding.

It may be noted that the different number of bits of $\mathcal{H}_o$ are used for selection of row objects and attributes to eliminate any common pattern in their selection. Further, for selecting bit position to conceal a watermark bit within an attribute of the selected object type attribute; S_f is concatenated with $\mathcal{K}_s$ to get a value whose mod with α yields position where a watermark bit is concealed (line 7).

Different watermark bit is selected for every earmarked row object. However, in a particular row object, same watermark bit *WIndex*(r) is concealed repeatedly within different attributes of the selected object type attribute. This adds redundancy resulting in enhancing robustness of $\mathcal{W}_o$.

B. **Insert Watermark**: We embed a watermark in independent object type attributes to yield watermarked database $\mathcal{O}_w$. The owner specifies the attribute set $A_\beta$comprising of attributes that can tolerate small amount of perturbations without compromising usability constraints. Usability constraints are pre-defined for every attribute by owner of a database depending upon its application.

We conceal watermark into fractional part of numeric type attributes present within the selected object. We refer to them as float attributes. However, the technique can be easily extended for varied multimedia data types supported by ORDB such as image, audio, video, etc. by following any of the existing techniques [4, 6, 22, 31].

We show recursive sub-routine in Fig. 4 that embeds a watermark in fractional part of the numeric attribute to introduce minimum distortions. The row objects whose $W_Status(r)$ is 1, participate in watermarking process. Within a selected row object, attribute A_{AIndex} is chosen among A_{β}. All the attributes within a selected object type A_{AIndex} are scanned. For every selected object type attribute, the watermark bit selected $\mathcal{W}[WIndex(\ r)]$ is embedded using pseudo-code in Fig. 4. Selected attribute A_{AIndex} of float type is passed as a parameter to *Embed_W(.)*. Each watermark bit is embedded repeatedly within a selected attribute by recursively calling *Embed_W(.)* subroutine.

It may be noted that use of OIDs to select secret target locations for embedding watermark and Object type attributes for embedding watermark enhances the security and robustness of the proposed scheme.

4.3 Prepare watermark for tamper detection

The third step of *WORD* is to prepare signature $\mathcal{W}_s$ of the watermarked database for tamper detection as depicted in Fig. 2. We take the signature of database $\mathcal{O}_{\mathrm{w}}$, which is concealed with the ownership watermark to create a fragile watermark. The process of creating $\mathcal{W}_s$ comprises of following sub-processes:

A. **Partition Database:** All the row objects in $\mathcal{O}_{\mathrm{w}}$ are divided into N_p virtual partitions. We define Partition set $P = \{P_0, \ldots, P_{I_d}, \ldots, P_{N_p-1}\}$ as set consisting of $N_p$virtual partitions such that for any two partitions $P_i \cap P_j = \{\varnothing\}$, *if* $i \neq j$. A unique partition number is assigned to every row object, $r\ \varepsilon\ \mathcal{O}_{\mathrm{w}}$ using Eq. 3.

$$I_d(r) = mod\big(\mathrm{OID}(r), N_p\big) + 1 \tag{3}$$

Where, *OID*(*r*) is *OID* of a row object and N_p is the number of partitions.

Embed_W ($\mathcal{A}$)

1. **For each** attribute $\mathbb{A} \in \mathcal{A}$
2. **If** datatype($\mathbb{A}$) is float
3. Embed $\mathbb{W}_p[WIndex]$ into $\mathbb{A}[BPos]$
4. **ElseIf** $datatype(\mathbb{A})$ is object type
5. Call *Embed_W(*$\mathbb{A}$*)*
6. **End If** (Line 2).
7. **End For** (line 1)

Fig. 4 Pseudo-code to embed watermark into selected attribute

B. **Create Watermark W_s**: Figure 5 illustrates the process of creating a signature watermark for each of the N_p virtual partitions by utilizing all the row objects assigned to the partition. It returns the composite watermark which is registered with a TTP.

While preparing for partition watermark, a secure row object's watermark hash $\mathcal{H}_w(r)$ that acts as a signature of a row object is generated by picking up the attribute values (line 3). Attributes in ORDB are of varied data type. Each attribute of a row object is processed and converted to real number. These values are used to generate the watermark hash $\mathcal{H}_w(r)$ of each row object as shown in Eq. 4.

$$\mathcal{H}_w(r) = Hash\left(Aval(A_1(r)) \|\ldots Aval(A_{AIndex}(r))\ldots\|Aval(A_{N_a}(r))\right) \tag{4}$$

Where, $Aval(A_i(r))$ denote the value of an attribute A_i for the row object r, such that: $1 \le i \le N_a$. Next, each row object's watermark $w_P[r][1:N_{wp}]$ is obtained by taking modulus of $\mathcal{H}_w(r)$ with $2^{N_{wp}}$ and converting the decimal into binary form (line 4). N_{wp} is the length of watermark in each partition decided by the owner. Finally, the partition watermark $\mathcal{W}_P$ of length N_{wp} is calculated by taking the *XOR* of the corresponding watermark bits for all row objects assigned to the partition (line 6):

$$\mathcal{W}_P\left[1:N_{wp}\right] = XOR\left(w_P[r_1]\left[1:N_{wp}\right], w_p[r_2]\left[1:N_{wp}\right]\ldots.\right) \tag{5}$$

Where, $w_P[r][1:N_{wp}]$ denotes the N_{wp} bits of the watermark created for the row object $r$of partition P. Watermarks of all the partitions are combined to get final watermark as $\mathcal{W}_s = \mathcal{W}_1 \|\ldots\|\mathcal{W}_P\|\ldots\ldots\|\mathcal{W}_{Np}$ comprising of $N_{wp} * N_p$ bits.

4.4 Register with TTP

In the fourth step, Owner registers the watermark $\mathcal{W}_s$ created in above step with third party for tamper detection. TTP provides transparent security to the process of watermarking.

Create_W (.)

Input: $\mathcal{O}_w, P, N_{wp}$
Output: $\mathcal{W}_s$

1. **For** every partition $P \varepsilon \mathcal{O}_w$
2. **For** every row-object r ε P
3. Calculate $\mathcal{H}_w(r)$ of r using equation 4
4. Calculate row object watermark: $w_P[r]\left[1:N_{wp}\right] = dec2bin\left(mod(\mathcal{H}_w(r), 2^{N_{wp}})\right)$
5. **End for** (Line 2)
6. Calculate partition watermark $\mathcal{W}_P\left[1:N_{wp}\right]$ using equation 5
7. **End for** (Line 1)
8. $\mathcal{W}_s = \mathcal{W}_1 || \ldots || \mathcal{W}_P || \ldots\ldots || \mathcal{W}_{Np}$

Fig. 5 Pseudo-code for creating the watermark $\mathcal{W}_s$

4.5 Generate decision

This is the last step of the proposed watermarking model as depicted in Fig. 2. A database may be used for varied purposes depending on its application domain. Thus, the need of database protection varies accordingly. Depending on whether ownership rights or integrity of the database is violated, this phase generates the decision. We now discuss how a decision is generated in each of following cases.

A. **Ownership Proof:** An attacker might try to destroy the watermark embedded in database and then claim ownership over a pirated copy. Generate decision process is responsible for proving ownership rights over suspected database $\mathcal{O}'$. Watermark $\mathcal{W}_x$ is successfully extracted from the suspected database $\mathcal{O}'$ using pseudo-code inscribed in Fig. 6. The process starts by calculating target locations (line 4-7). It then extracts the watermark bits by calling recursive subroutine *Extract_W(.)*inscribed in Fig. 7.

We apply majority voting to decide the watermark bits of $\mathcal{W}_x$ (line 11-18). Majority voting is a decision rule which selects one value which has occurred maximum number of times. For each watermark bit position b, we count the number of 1 s (*Count*[b][1]) and 0 s (*Count*[b][0]) extracted. If number of ones is greater than zeros for a particular bit b, then we assign $\mathcal{W}_x[b]$ = 1 else $\mathcal{W}_x[b] = 0$. In case both are equal we consider that bit as invalid.

The bits of the watermark extracted $\mathcal{W}_x$ from a suspected database are matched with the corresponding bits of the original watermark $\mathcal{W}$. The original watermark $\mathcal{W}$ is either stored or can be re-generated using process described in Section 4.1. If the number of matches N_m is

Extract_Watermark (.)

Input: $\mathcal{O}', \mathcal{H}_o(r), \gamma, \beta, \alpha, \mathcal{K}_s, N_w$
Output: Extracted watermark $\mathcal{W}_x$

1. Assign $Count[1{:}N_w][0] = 0, Count[1{:}N_w][1] = 0$
2. **For** every row-object $r \; \varepsilon \; \mathcal{O}'$
3. $S_f = \mathcal{H}_o(r)[1{:}20]$, $S_e = \mathcal{H}_o(r)[1{:}30]$
4. **If** $(S_f mod \; 1/\gamma) == 0)$
5. $AIndex = S_e mod \; \beta$
6. $BPos = mod\left((S_f || \mathcal{K}_s), \alpha\right)$
7. $WIndex(r) \; = mod(S_e, N_w)$
8. Call $Extract_W(A_{AIndex})$ described in figure 3.7
9. **End If**
10. **End For** (line 2)
11. **For** b=1 to N_w
12. **If** $(Count[b][1] > Count[b][0])$
13. $\mathcal{W}_x[b] = 1$
14. **Else If** $(Count[b][1] < Count[b][0])$
15. $\mathcal{W}_x[b] = 0$
16. **Else**
17. $\mathcal{W}_x[b]$ not detected.
18. **End If** (line 12).
19. **End For** (line 11)

Fig. 6 Pseudo-code for extracting the ownership watermark $\mathcal{W}_x$

Extract_W (𝒜)

1. **For each** attribute $\mathbb{A} \varepsilon \mathcal{A}$
2. **If** datatype($\mathbb{A}$) is float
3. b=$\mathbb{A}[BPos]$
4. $Count[WIndex][b]$ += 1
5. **Else If** datatype($\mathbb{A}$) is object type
6. Call *Extract_W($\mathbb{A}$)*
7. **End If** (Line 2).
8. **End For** (line 1)

Fig. 7 Recursive algorithm to extract the watermark from selected attribute

very large we suspect piracy. To assess such quantities, a threshold τ_w is taken. We define τ_w as the threshold value that decides the database piracy. It tells the permissible fraction of matches of bits of $\mathcal{W}_x$ with $\mathcal{W}$ to claim ownership over the disputed database. This value is decided by the owner depending on the sensitivity of data. Equation 6 shows the condition to detect the piracy.

$$If\,(N_m \geq \tau_w * N_w),\, then\ database\ is\ pirated. \tag{6}$$

It may be noted that the proposed watermark extraction technique is blind as it does not require the original database for extracting the watermark from $\mathcal{O}'$.

B. **Tamper Detection:** Whenever a database is transmitted across network, receiver of the database checks for any perturbations introduced during network transmission. In order to detect any attempt to tamper the contents of the database, we regenerate the watermark $\mathcal{W}_g$ from the suspected database $\mathcal{O}'$. We partition the suspected database using Eq. 3 to get P', and then invoke subroutine *Create_W(.)* inscribed in Fig. 5 to get $\mathcal{W}_g$. Inputs to the subroutine *Create_W(.)* are $\mathcal{O}'$ and P'. The watermark is created by using attribute values of all row objects in the suspected database. Hence, any modification in the database will result in the creation of a different watermark. Owner retrieves the registered watermark $\mathcal{W}_s$ from the TTP. Both the watermarks are then compared. If $\mathcal{W}_g \neq \mathcal{W}_s$; database has been tampered with. For every partition, different watermark is created and concatenated to yield final one. This provides localization up to partition level to our tamper detection scheme.

It may be noted that tamper detection process does not require any of secret parameters used in ownership proof; thereby maintaining security of ownership process.

5 Security analysis

Assume that Alice is the owner of the database $\mathcal{O}$ and has embedded watermark $\mathcal{W}$ to generate a watermarked data set $\mathcal{O}_w$. Mallory is an attacker who has no access to the original database $\mathcal{O}$ and does not know any of the secret parameters: object identifier *OID*, secret key K_s, number of partitions N_p, fraction of watermarked row objects γ and the number of candidate attributes β. Mallory may try to destroy the watermark to claim the database to be hers. Under

such circumstances, we justify the security of our model with respect to the following parameters:

i. **Probability to identify watermarked positions**

We describe *WORD* as secure mechanism making it extremely difficult for Mallory to find the positions of embedded bits. *WORD* embeds watermark into γ fraction of the total N_t row objects. Further, it selects one out of β candidate attributes for watermarking within selected row objects. Obviously, without knowing the secret parameters γ and β, an attacker will find it impossible to correctly identify the watermarked row objects. We now calculate the probability of correctly identifying watermarked object types.

Probability of correctly choosing single watermarked row object is:

$$P_{r1} = \gamma * N_t / N_t \tag{7}$$

Probability to correctly identify all the $\gamma * N_t$watermarked row objects is given as:

$$\begin{aligned} P_r &= \frac{\gamma * N_t}{N_t} * \frac{(\gamma * N_t) - 1}{(N_t - 1)} * \frac{(\gamma * N_t) - 2}{(N_t - 2)} * \ldots\ldots\ldots * \frac{1}{(N_t - \gamma * N_t + 1)} \\ &= \left(\frac{1}{{}^{\mathrm{N}}_{\gamma * Nt}\mathrm{t}} C^{\gamma * N_t} \right) \end{aligned} \tag{8}$$

Where, γ is fraction of tuples selected for watermarking, N_t is total number of row objects and C is combination of N_t row objects taken $\gamma * N_t$ at a time without repetition. Without knowing A_β, Mallory will select A_F as a set of all the object type that can be in candidate attributes set; such that $A_\beta \subseteq A_F$. An earmarked object type attribute can be chosen in $1/\beta_F$ ways.

Thus, probability of correctly selecting an object type per row object is given as:

$$P_{or} = P_{r1} * \frac{1}{\beta_F} = \frac{\gamma * N_t}{N_t * \beta_F} \tag{9}$$

Probability to correctly identify targeted locations with respect to watermarked object type of all the $\gamma * N_t$ selected row objects is given as:

$$\begin{aligned} P_l &= \frac{\gamma * N_t}{N_t * \beta_F} * \frac{(\gamma * N_t) - 1}{(N_t - 1) * \beta_F} * \frac{(\gamma * N_t) - 2}{(N_t - 2) * \beta_F} * \ldots\ldots\ldots * \frac{1}{(N_t - \gamma * N_t + 1) * \beta_F} \\ &= \left(\frac{1}{{}^{\mathrm{N}}_{\gamma * Nt}\mathrm{t}} C^{\gamma * N_t} \right) * \left(\frac{1}{\beta_F} \right)^{\gamma * N_t} \end{aligned} \tag{10}$$

Where, C is combination of N_t row objects taken $\gamma * N_t$ at a time without repetition. Substituting $\gamma = 1/3$, $\beta_F = 10$, $N_t = 6000$ in Eq. 10, we get $P_l = 0.22\ e^{-3656}$ which is negligible. Hence, it is impossible for the Mallory to correctly identify the watermarked positions. This probability further decrease, depending on the number of float attributes each watermarked object type contains. Tables 4 and 5 show the effect of P_l at different values of β_F. As a number of candidate attributes increases, probability decreases further. Similarly, the probability is inversely proportional to γ *and* N_t, as it decreases with the increase in γ *and* N_t.

Table 4 Effect of β_F *on* P_l $(*e^{-2000})$ at different values of γ

	$\gamma = 1/2$	$\gamma = 1/3$	$\gamma = 1/5$
$\beta_F = 5$	$0.78e^{-1901}$	$0.25e^{-1054}$	$0.16e^{-140}$
$\beta_F = 10$	$0.64e^{-2804}$	$0.22e^{-1656}$	$0.90e^{-502}$
$\beta_F = 15$	$0.34e^{-3332}$	$0.14e^{-2008}$	$0.45e^{-713}$
$\beta_F = 20$	$0.52e^{-3707}$	$0.19e^{-2258}$	$0.53e^{-864}$

ii. False Hit Rate

P_{FH} is the probability of extracting a valid watermark from a non-watermarked database. Our robustness analysis is grounded upon the approach given in [18]. Let a watermark bit b_i be embedded N_r times in a partition. For simplicity, we calculate N_r assuming one watermark bit per row-object as $N_r = \left\lfloor \frac{\gamma * N_t}{N_w} \right\rfloor$. Each extracted character b_i^* from a non-watermarked relation has same probability of 1/2 to match or not match the original embedded watermark bit in the database. The final watermark bit is decided based on majority voting. We define P_{bit} as the probability of correctly extracting one watermark bit after applying majority voting by sheer chance. This is equivalent to saying that P_{bit} is the probability that at-least one more than half of N_r bits can be detected from non-watermarked relation by a sheer chance. We now calculate P_{bit} using binomial distribution considering N_r independent trials [23] as:

$$P_{bit} = \sum_{j=N_r/2+1}^{N_r} b\left(j; N_r, \frac{1}{2}\right) = B((0.5 * N_r + 1); N_r, 0.5) \tag{11}$$

Where, $b(j; n, p)$ is the probability of obtaining j successes in n Bernoulli trials with probability p for success and $1-p$ for failure and is given in Eq. 12.

$$b(k; n, p) = \binom{n}{k} C) * (p^k) * (1-p)^{n-k} \tag{12}$$

We use $B(k; n, p)$ referred as cumulative binomial probability representing the probability of obtaining at-least k successes from n Bernoulli's trials [23]; given as:

$$B(k; n, p) = \sum_{i=k}^{n} b(i; n, p) \tag{13}$$

For a watermark of length N_w, we calculate false hit rate P_{FH} referring Eq. 6 as:

$$P_{FH} = B(\tau_w * N_w; N_w, P_{bit}) \tag{14}$$

Where, τ_w is watermark threshold, N_w is length of watermark in bits and P_{bit} as given in Eq. 11. We took $N_t = 6000$, $N_w = 50$ and plotted the graph by varying τ_w. Figure 8 illustrates that false hit rate is monotonically decreasing with increase in threshold τ_w. False hit rate P_{FH} is low for

Table 5 Effect of β_F *on* $P_l(*e^{-2000})$ at different values of N_t

	$N_t = 3000$	$N_t = 6000$	$N_t = 9000$
$\beta_F = 05$	$0.34e^{474}$	$0.25e^{-1054}$	$0.17e^{-2582}$
$\beta_F = 10$	$0.32e^{-827}$	$0.22e^{-1656}$	$0.37e^{-3485}$
$\beta_F = 15$	$0.26e^{-003}$	$0.14e^{-2008}$	$0.73e^{-4014}$
$\beta_F = 20$	$0.29e^{-128}$	$0.19e^{-2258}$	$0.12e^{-4388}$

the nominal value of $\tau_w = 0.6$, which means that our method is robust and secure. We observed the effect of γ on P_{FH} and plotted the values in same graph. It is revealed that P_{FH} increases with increase in γ. As more number of row objects are earmarked, P_{FH} increases.

iii. **False Miss Rate**

We define false miss rate, P_{FM}as the probability of not detecting the watermark from a watermarked database. Mallory will try to destroy the watermark through various attacks. If Mallory successfully deletes $(\tau_w * N_w + 1)$ watermark bits out of N_w bits, watermark is destroyed. We define P_{FM} as given in Eq. 15.

$$P_{FM} = \prod_{i=1}^{\tau_w * N_w + 1} Prob(i) \tag{15}$$

Where, $Prob(i)$ is probability of successfully destroying i^{th} watermark bit. We now analyse the false miss rate against Subset deletion, addition and alteration attacks.

A. **Subset Deletion Attack:** In this attack, the Mallory deletes $\oint$ row objects from the watermarked database $\mathcal{O}_w$ with the intention of deleting earmarked row objects. Each watermark bit is embedded N_r times in the database. To destroy i^{th} watermark bit; $\oint$ random deletions must include N_r marked row objects containing i^{th} watermark bit. Probability that i^{th} watermark bit is completely destroyed with $\oint$random deletions is given by hyper-geometric distribution [10] as:

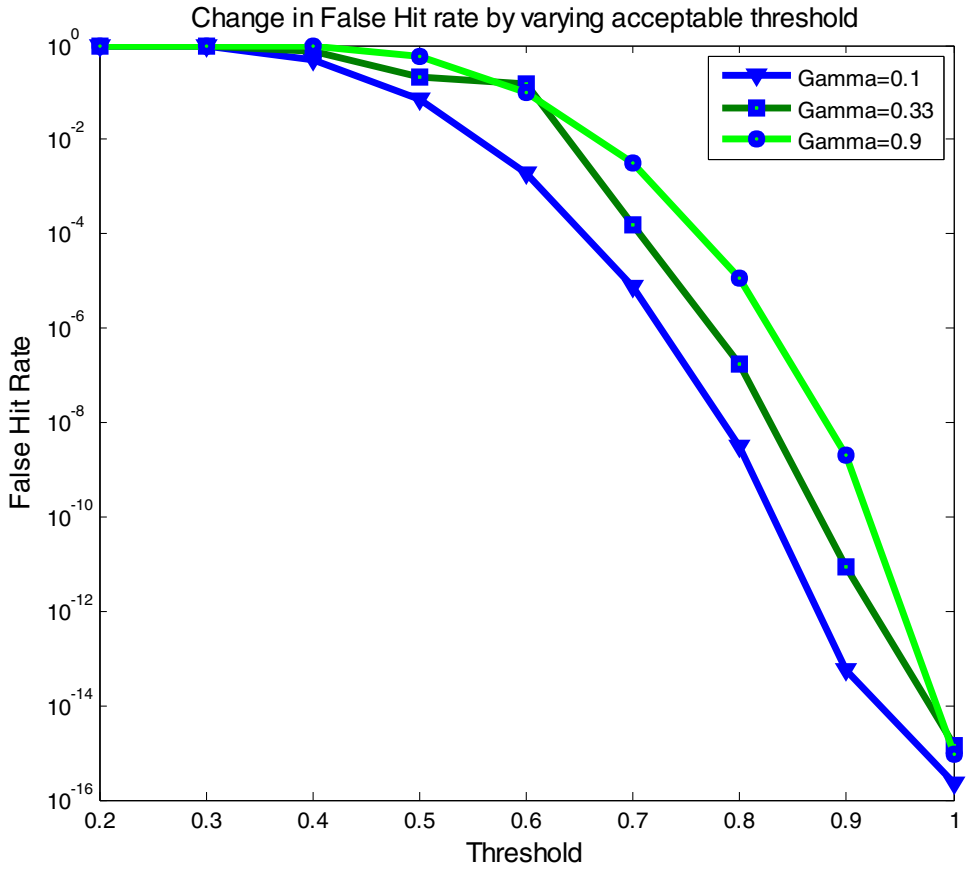


Fig. 8 Change in false hit rate by varying watermark threshold τ_w *and* γ

$$Prob(i) = \left(\frac{\binom{\mathrm{Nr}}{Nr} C) * \binom{\mathrm{Nt-Nr}}{\phi\oint - Nr} C^{\phi\oint - N_r}}{\overset{\mathrm{N}}{\oint} \mathrm{t}} C^{\oint} \right) \tag{16}$$

Suppose a population consists of N items, k of which are successes. And a random sample drawn from that population consists of n items, x of which are successes. Then the hypergeometric probability is:

$$h(x; N, n, k) = \left(\binom{k}{x} C * \frac{\overset{\mathrm{N-k}}{n-x} C^{n-x}}{\overset{N}{n} C)} \right. \tag{17}$$

Taking $\oint = 1000$, $N_t = 6000$, $\gamma = 1/3$, $\tau_w = 0.6$, $N_w = 50$, we calculated $P_{FM(Subset\ Deletion)} = 3.2\ e^{-413}$ by substituting Eq. 16 in 15. Thus P_{FM} is negligible and close to zero for subset deletion attack. We plotted the graph in Fig. 9 of $Prob(i)$ by varying value of $\oint$ at different values of γ.The results show that with the increase of $\oint$, probability increases. However, this value is very less even for $\oint = 4000$. If an attacker deletes beyond this, database becomes useless. It is also observed that as γ increases, probability decreases. At $\gamma = 1/3$, we get very low probability making our scheme secure. As more number of row objects is involved in watermark insertion, $P_{FM(Subset\ Deletion)}$ further decreases.

B. **Subset Addition Attack:** In this attack, Mallory tries to fudge the embedded watermarks by adding $\oint$ fake row objects to the watermarked database $\mathcal{O}_w$. Recall that she has no knowledge of any of the secret parameters. The additional row objects do not contain watermark bits but insert noise. Let us consider the worst-case scenario when any row object being added is selected and contains all inverted watermark bits. In order to completely obliterate the watermark bit, Mallory has to add minimum of earmarked row objects corresponding to i^{th} watermark bit i.e. (N_r). Probability that i^{th} watermark bit is completely destroyed with $\oint$ random additions of row objects is same as calculated above in Eq. 16; which is negligible. Hence, the proposed technique is robust against this attack.

C. **Subset Alteration Attack:** In this attack, Mallory tries to modify $\oint$ row objects randomly to alter the embedded watermark. Let us consider the worst-case scenario when any row object after being altered contains all inverted watermark bits. In order to completely obliterate the watermark bit, Mallory has to alter one more than half of earmarked row objects corresponding to i^{th} watermark bit i.e. $((N_r/2) + 1)$. Probability that i^{th} watermark bit is completely destroyed with $\oint$ random alterations of row objects is given by hypergeometric distribution [10] as:

$$Prob(i) = \left(\frac{\binom{(\mathrm{Nr})}{Nr/2+1} C^{(N_r/2+1)}) * \binom{\mathrm{Nt-Nr}}{\phi\oint - Nr/2-1} C^{\phi\oint - N_r/2-1}}{\overset{\mathrm{N}}{\oint} \mathrm{t}} C^{\oint} \right) \tag{18}$$

Taking $\oint = 1000$, $N_t = 6000$, $\gamma = 1/3$, $\tau_w = 0.6$, $N_w = 50$, wecalculated $P_{FM(Subset\ Alteration)} = 1.2e^{-88}$ by substituting Eq. 18 in 15, thus P_{FM} is negligible. We plotted the graph of $Prob(i)$ by varying value of $\oint$ at different values of γ as shown in Fig. 10. The results show that with

the increase of $\oint$, probability increases significantly till $\oint = 3000$. Beyond this, there is marginally decrease in P_{FM}. Further, it is observed that as γ increases, probability decreases. At $\gamma = 1/3$, we get low P_{FM} making our scheme robust against this attack.

iv. **Integrity Analysis**

Another objective of WORD is to check the integrity of a suspected database. For tamper detection, watermark should be fragile to detect even a slightest change made in a database. In the following analysis, we assume single attribute alterations per row object and analyse the probability of not detecting any tampering after different attacks.

A. **Subset Alteration:** The procedure creates a signature of the entire database; any modification to the database is bound to alter its signature. Therefore, if the attacker tries to make any changes to earmark or non-earmark attributes in the database, then on recreating the watermark from the suspected database will always produce a different watermark from the one registered and hence temperedness will be detected.

B. **Subset Addition:** When an attacker adds a random row object, its N_w bit watermark is *xored* with watermarks of other row objects. There is only one combination, i.e. all zeros that will retain the original watermark. Thus in this case, the probability of not detecting tampering through addition of a single row object is defined as:

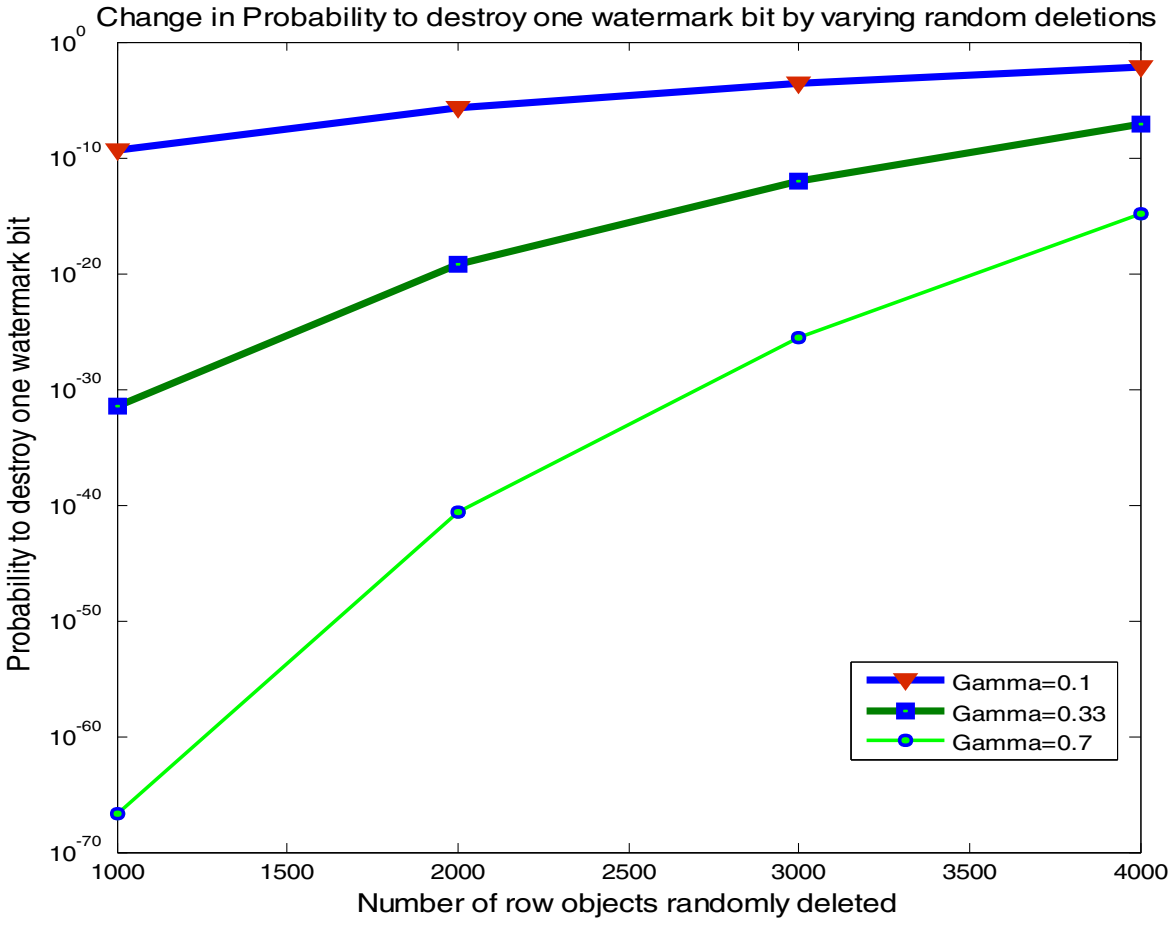


Fig. 9 Graph showing change in probability to destroy a watermark bit against random deletions by varying γ

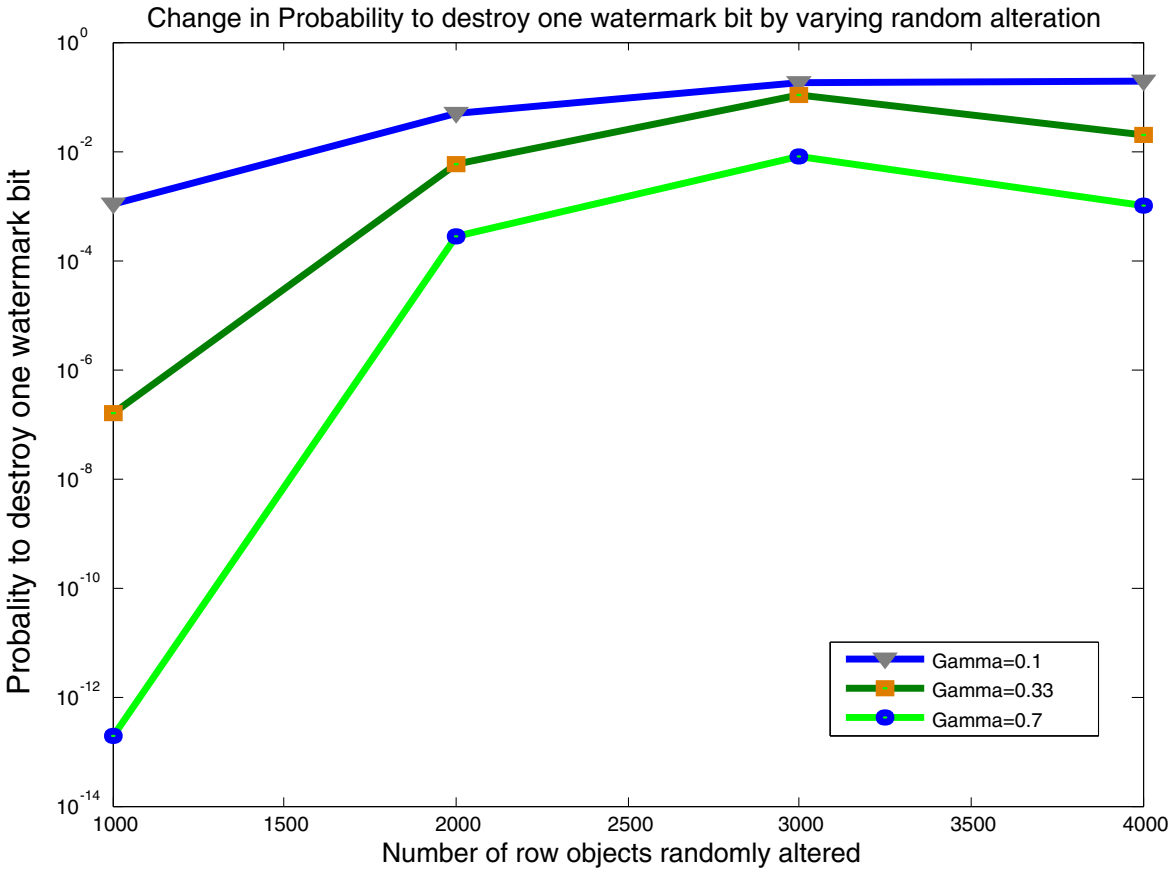


Fig. 10 Graph showing change in probability to destroy a watermark bit against random alterations by varying γ

$$P_{Addition} = \frac{1}{2^{N_w}} \tag{19}$$

For $N_w = 50$, $P_{Addition} = 0.8e^{-15}$ which is negligible. As the length of watermark and number of row objects added increases, this probability further decreases.

C. **Subset Deletion:** In case of deletion of a row object, the tampering will not be detected only if the deleted row object's watermark contained all zeros. This is produced with a probability = $1/\left(2^{N_w}\right)$. Furthermore, probability that the attacker deleted this particular row object = *1/Total row objects* =$1/N_t$.Thus, probability of not detecting tampering by subset deletion is defined as:

$$P_{Deletion} = \frac{1}{2^{N_w}} * \frac{1}{N_t} \tag{20}$$

For $N_w = 50$, $P_{Addition} = 0.15e^{-17}$. As the length of watermark and/or number of row objects increases this probability further decreases.

6 Implementation of WORD

We have modeled the system using pattern designs to incorporate flexibility and scalability. Since design patterns give a higher-level perspective on the problem and on the process of design and object orientation, *WORD* is developed using strategy pattern. This illustrates

"design-with-change-in-mind" approach [29]. Figure 11 shows the class diagram of *WORD*. In consonance with the philosophy of design patterns, each of the classes shown has a pre-designated responsibility. Association between classes is represented using different arrows following Unified Modelling Language (UML) notations. Inheritance is represented using arrow head and aggregation using open diamond. Dashed line with an arrow depicts dependency. Data members of classes are made protected or private while member functions are made public.

The *Watermark _ Manager* (*WM*) is a singleton class that initiates the watermarking process by accepting watermarking requests and initializing a set of secret parameters required for watermarking. It collaborates with the *Watermark _ Embedder*(*WE*) and the *Security _ Manager*(*SM*) classes to realize the core functions involved in watermarking process. The class *WE* specialize into *ObjectType _ WE* class which in turn is extended by *Numeric _ WE, Image _ WE, Sound _ WE* and *Video _ WE* classes. These classes are responsible for implementing specific strategies for preparing and embedding bits of ownership watermark for various data types. The strategy pattern allows one to incorporate any new kind of embedding approach as and when the need arises. *WE* uses *Watermark _ Creation* (*WC*) class for preparing $\mathcal{W}_o$ and $\mathcal{W}_s$.*WC* has class *Partition _ Assigner*(*PA*) that creates N_p partitions of a database.

Security _ Manager (*SM*) class discusses two important security goals; ownership proof and tamper detection. Watermarking techniques can also address other security concerns such as data provenance [27], fingerprinting [21] and their combination. However, the security concern for a database can change at different points of time. It is therefore, judicious to adopt

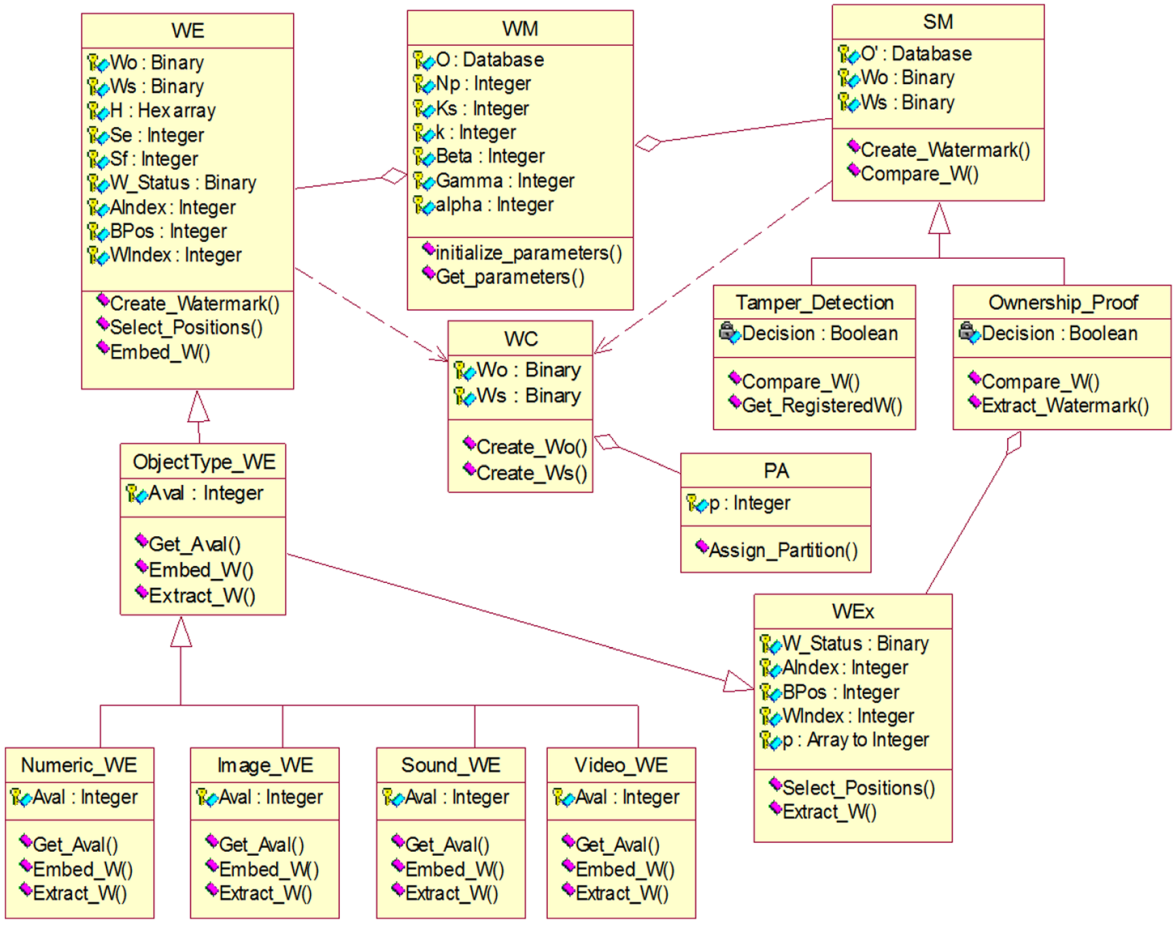


Fig. 11 Class diagram showing implementation details of *WORD*

a strategy pattern that abstracts security as a conceptual concern and uses specialized methods in inherited classes for addressing specialized security issues. This allows room for adding new security functionality. In the Fig. 11, we have shown *Tamper _ Detection* and *Ownership _ Proof*as specialized strategies of *SM* due to focus of our work on these issues. *Ownership _ Proof* has *Watermark _ Extractor* (*WEx*) that extracts the embedded watermark from specialized abstract class *ObjectType _ WE*. Depending on an attribute data type used for embedding, the appropriate data type class overrides *Extract _ W*(.) function of *ObjectType _ WE* class. For the ease of understanding, we have included only the important data members and member functions in their respective classes.

7 Experimental results and analysis

We conducted our experiments on an Intel Core™ i7 2.30 GHz system in MATLAB 7.8.0. We have applied our technique on the generated synthetic object oriented database developed using PostgreSQL 9.2.1. PostgreSQL is an open-source object relational database management system. Data set comprises of 50,000 row objects. The various experiments performed on this data-set to test its robustness and integrity is discussed in following subsections.

7.1 Robustness analysis

When a database is freely distributed over a network, an attacker Mallory could perform different types of attacks in an attempt to perturb or even delete the embedded watermark. It may be noted that Mallory is faced with the dilemma of trying to destroy the watermark without affecting the usability of the data. We have assumed that all the perturbations made by owner to embed the watermark or done by an attacker to destroy the watermark are within usability constraints. We now simulate the effects of different types of attacks on database to demonstrate its robustness.

A. **Subset Deletion Attack:** In this attack, the Mallory deletes ϕ row objects from the watermarked database $\mathcal{O}_w$ with the intention of deleting earmarked row objects. The experimental results shown in Fig. 12 reveal that even on 90% deletion of row objects, the watermark is extracted with 100% accuracy. This is possible owing to increased redundancy of watermark bits and majority voting mechanism. This verifies the results of theoretical analyses shown in Section 5 proving resilience of our technique against this attack.
B. **Subset Addition Attack:** In this attack, Mallory tries to fudge the embedded watermarks by adding ϕ random row objects to the watermarked database $\mathcal{O}_w$. Assume that she has no knowledge of any of the secret parameters. The additional row objects do not contain watermark bits but insert noise. However, each added row object is equally likely to get selected and contain same or reverse watermark bits. Till the number of row objects added N_n is less than total row objects N_t, the correct watermarking bit is extracted. If ($N_n > N_t$), then extracted bit may differ. This result in adding more than 100% noise in dataset; thereby destroying its usability. Experimental results shown in Fig. 12 reveals that even on adding 100% new row objects; we are able to extract watermark thus verifying the results of theoretical analyses discussed previously in Section 5.

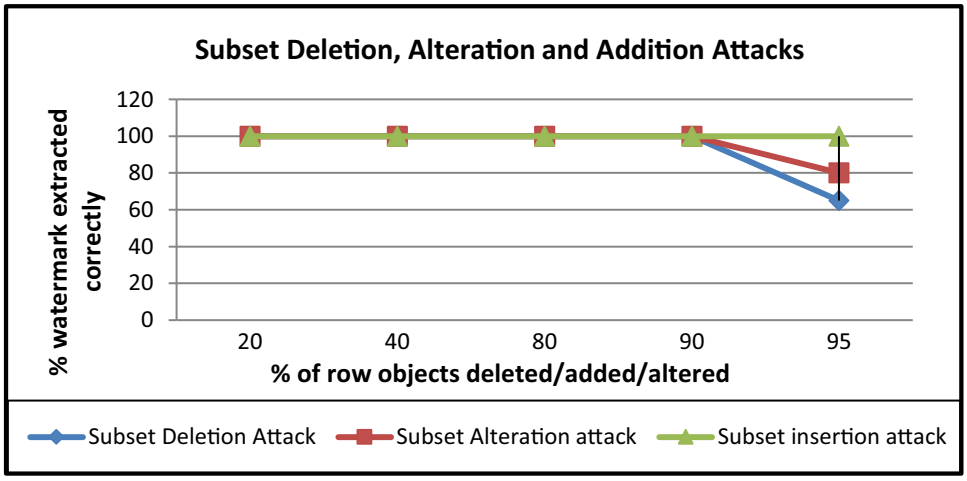


Fig. 12 Percentage of watermark extracted correctly by subset deletion, insertion and alteration attack

C. **Subset Alteration Attack:** In this attack, Mallory tries to modify ϕ row objects randomly to alter the embedded watermark. In subset alteration attack, we assume that attacker will alter few attributes per row object randomly. The experimental results in Fig. 12 verify the robustness of the proposed scheme. On 90% alteration, the watermarking system *WORD* could recover the complete watermark by tapping the inherent redundancy within a row object using majority voting.

D. **Row object and Attribute Reordering:** In this attack, the Mallory reorders the position of the row objects/attributes in the database with an intention of destroying the embedded watermark. To counter attribute reordering, we suggest attributes to be sorted first and then their indices are selected for concealing watermark. Experimental results reveal 100% resilience against this attack. In our technique, position of a row object in a database holds no significance while creating, embedding or extracting a watermark from a database. We use the *OID* hash which cannot be altered by simply repositioning a row object or any attribute in it. Thus, our technique shows robustness towards this attack.

E. **Attribute Deletion Attack:** In this attack, Mallory deletes the attributes of a database hoping to delete attributes with watermarks embedded. We embed a watermark in randomly chosen attribute from the list of candidate attributes, which is different for each row object. Graph in Fig. 13 shows percentage of watermark extracted correctly on deleting the candidate attributes. On deleting nearly 50% of candidate attributes, the watermark is accurately extracted. On further deletion, it degrades gradually.

F. **Watermark Synchronisation Error:** In this attack, the watermark is shifted by a few positions relative to its originally embedded position. In our proposed technique, watermark index is calculated using *OID*, hence is dependent on row object. Every selected row object will determine the watermark bit index that is embedded in it. Even if certain earmarked row objects are deleted, this will not affect the watermark bit positions.

7.2 Integrity analysis

We performed experiments on the subset attacks. Figure 14 shows the percentage of a recreated watermark match with their corresponding original registered one when

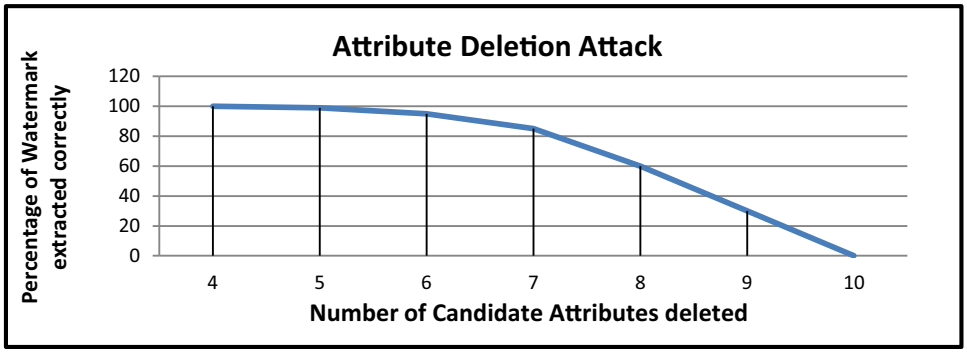


Fig. 13 Percentage of watermark extracted in attribute deletion attack

subjected to subset addition, deletion and modification attacks. It is observed that with a 5% addition, deletion or alteration of the database, there is as much as 50%mismatch. This indicates that even with the slightest modification, the recreated watermark changes significantly, thus showing unambiguously that database has indeed been tampered with.

7.3 Imperceptibility

To analyse the amount of distortions introduced after watermark embedding, we use mean and variance as usability constraints. Taking $N_t = 6000$, $\gamma = 1/3$; the change in the mean and variance are calculated for different attributes and are recorded in Table 6. The results indicate that the amount of errors introduced is very low. Hence, the watermark can be safely embedded without significant distortion to the database.

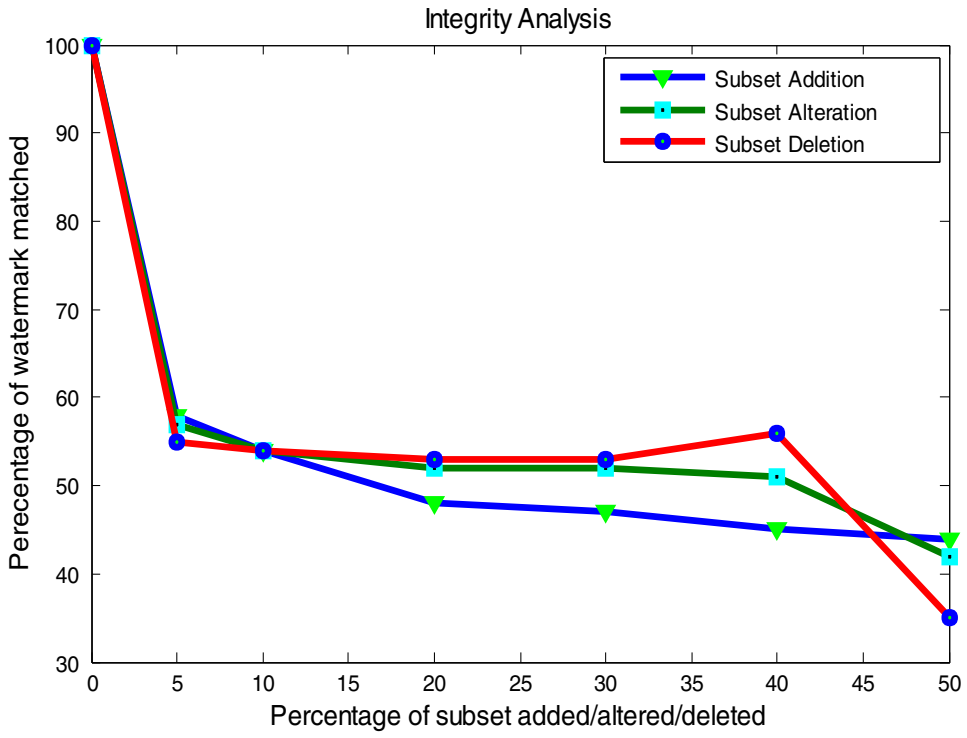


Fig. 14 Integrity analysis against subset addition, alteration and deletion attack

Table 6 Effect of watermarking on mean and variance of data

Attribute type	% Change in mean	%Change in variance
A1	1.23	0.70
A2	0.28	1.2
A3	1.37	0.20
A4	0.58	0.70

8 Conclusion and future work

WORD provides a generic framework for protection of ORDBs. It is robust in its approach to extract the embedded watermark and fragile in its approach while re-creating the watermark from a suspected database. The sequential applications of two different watermarks using the services of a trusted third party provides ownership proof and tamper detection. Watermark is concealed within object type attribute using OID of each row object. This increases the security level, makes the watermarking technique primary key independent and gives us more bandwidth to accommodate watermark bits and hence results in increase in robustness. To deal with changing security aspects with time, strategy pattern can be adopted while implementing the watermarking model *WORD*. The use of strategy pattern designs incorporates flexibility and scalability in our model.

Prior research mainly focussed on RDBs. We have designed a generic scheme for protection against tamper detection and ownership proof in ORDBs. The direction now points towards: (i) extending from ORDBs to OODBs and other emerging databases such as No-SQL databases flourishing in the web (ii) extending the scheme to address additional security concerns like finger-printing, data provenance etc.

Thus, there is a need to design suitable watermarking techniques targeting emerging web databases to provide protection against various security aspects.

References

1. Agarwal N, Singh AK, Singh PK (2019) Survey of robust and imperceptible watermarking. Multimed Tools Appl:1–31. https://doi.org/10.1007/s11042-018-7128-5
2. Agrawal R, Haas PJ, Kiernan J (2003) Watermarking relational data: framework, algorithms and analysis. VLDB J 12(2):157–169
3. Barrodale Computing Services Ltd. (BCS). http://www.barrodale.com/watertwapage/
4. Bashir T, Usman I, Albesher A, Almejalli KA, Naqvi SS (2019) GP based smart reversible watermarking of depth image based rendering for stereoscopic images. Multimed Tools Appl:1–20
5. Bhattacharya S, Cortesi A (2009) A distortion free watermark framework for relational databases. In: International conference on software and data technologies, vol 2, pp 229–234
6. Chen L, Zhao J (2018) Contourlet-based image and video watermarking robust to geometric attacks and compressions. Multimed Tools Appl 77(6):7187–7204
7. Farfoura ME, Horng SJ, Lai JL, Run RS, Chen RJ, Khan MK (2012) A blind reversible method for watermarking relational databases based on a time-stamping protocol. Expert Syst Appl 39(3):3185–3196
8. Halder R, Pal S, Cortesi A (2010) Watermarking techniques for relational databases: survey, classification and comparison. J Univ Comput Sci 16(21):3164–3190
9. Hoffer JA, Venkataraman R, Topi H (2010) Modern database management, 11th edn. Pearson, Boston
10. Hypergeometric distribution. http://stattrek.com/probability-distributions/hypergeometric.aspx
11. Kamel I (2009) A schema for protecting the integrity of databases. Comput Secur 28:698–709
12. Kerckhoffs A (1883) La cryptographiemilitaire. Journal des sciences militaries 9(1):5–83

13. Khan A, Husain SA (2013) A fragile zero watermarking scheme to detect and characterize malicious modifications in database relations. Sci World J 2013, Article ID 796726:1–16
14. Khanduja V (2017) Database watermarking, a technological protective measure: perspective, security analysis and future directions. JISA 37:38–49 ISSN: 2214-2126. Elsevier
15. Khanduja V, Chakraverty S (2018) Fragile watermarking of decision system using rough set theory. Arab J Sci Eng:1–13
16. Khanduja V, Chakraverty S, Verma OP, Singh N (2014) A scheme for robust biometric watermarking in web databases for ownership proof with identification. In: International conference on active media technology. Springer, Cham, pp 212–225
17. Khanduja V, Verma OP, Chakraverty S (2015) Watermarking relational databases using bacterial foraging algorithm. Multimed Tools Appl 74(3):813–839
18. Khanduja V, Chakraverty S, Verma OP (2016) "Ownership and tamper detection of relational data: framework, techniques and security analysis", published as the chapter in the book titled: embodying intelligence in multimedia data hiding at Science Gate Publishing, pp 21–36. https://doi.org/10.15579/gcsr.vol5. ISBN 978-618-81418-8-9 (print) - ISBN 978-618-81418-9-6 (e-book)
19. Khanduja V, Chakraverty S, Verma OP (2016) Enabling information recovery with ownership using robust multiple watermarks. JISA 29:80–92
20. Li Y, Guo H, Jajodia S (2004) Tamper detection and localization for categorical data using fragile watermarks. In: ACM workshop on digital rights management, pp 73–82
21. Li Y, Swarup V, Jajodia S (2005) Fingerprinting relational databases: schemes and specialties. IEEE Trans Dependable Secure Comput 2(1):34–45
22. Lie W, Chang L (2006) Robust and high quality time domain audio watermarking based on low frequency amplitude modification. IEEE Trans Multimedia 8(1):46–59. https://doi.org/10.1109/TMM.2005.861292
23. Lipschutz S, Schiller J (2005) Schaum'sOultilne of theory and problems of introduction to probability and statistics. McGraw Hill Education (India) Edition
24. Naz F, Khan A, Ahmed M, Khan MI, Din S, Ahmad A, Jeon G (2019) Watermarking as a service (WaaS) with anonymity. Multimed Tools Appl:1–25
25. Oracle Documentation. http://docs.oracle.com/cd/B19306_01/server.102/b14220/objects.htm
26. Ristic I (2014) SHA-1 Deprecation. Available: https://community.qualys.com/blogs/securitylabs/2014/09/09/sha1-deprecation-what-you-need-to-know.
27. Schäler M, Schulze S, Merkel R, Saake G, Dittmann J (2011) Reliable provenance information for multimedia data using invertible fragile watermarks, BNCOD, LNCS 7051, pp 3–17
28. Schneier B (2008) Applied cryptography, protocols, algorithms and source code in C, 2nd edn. Wiley-India
29. Shalloway A, Trott JR (2008) Design pattern explained, a new perspective on object-oriented design, second edn. Pearson education, Boston
30. Shehab M, Bertino E, Ghafoor A (2008) Watermarking relational databases using optimization-based techniques. IEEE Trans Knowl Data Eng 20(1):116–129
31. Singh SP, Bhatnagar G (2019) A simplified watermarking algorithm based on lifting wavelet transform. Multimed Tools Appl:1–22
32. Stonebraker M, Moore D (1995) Object relational DBMSs: the next great wave. Morgan Kaufmann Publishers Inc., San Francisco
33. Sun J, Cao Z, Hu Z (2008) Multiple watermarking relational databases using image. In: IEEE International Conference on Multimedia and Inf. Technology, pp 373–376
34. Tsai M, Tseng H, Lai C (2006) A database watermarking technique for tamper detection. In: International joint conference on information sciences. Atlantis Press. https://doi.org/10.2991/jcis.2006.206
35. Tufail H, Zafar K, Baig R (2018) Digital watermarking for relational database security using mRMR based binary bat algorithm. In: 2018 17th IEEE international conference on trust, security and privacy in computing and communications/12th IEEE international conference on big data science and engineering (TrustCom/BigDataSE). IEEE, pp 1948–1954
36. Wikipedia (2015) SHA-2. [Online]. Available: https://en.wikipedia.org/wiki/SHA-2

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

Dr. Vidhi Khanduja received her B.Tech degree in Information Technology from Guru Govind Singh Indraprashtha University, Delhi, India, M.E degree in Computer Technology and Applications from Delhi College of Engineering under University of Delhi, Delhi, India and Ph.D degree from University of Delhi, Delhi, India. Presently, she is working as Associate professor and Head of Department, Computer Engineering Department in SAL Institute of Technology and Engineering Research, SAL Education Campus, Ahmedabad, Gujarat. She had served as an Assistant Professor in Delhi College of Engineering Delhi, India and at Netaji Subhas Institute of Technology, University of Delhi, Delhi, India. Her research interests include Database protection, Information security, Evolutionary algorithms, Artificial Intelligence, Machine Learning, Super Image resolution.

Prof. Shampa Chakraverty is Professor in the Department of Computer Engineering in Netaji Subhas Institute of Technology, New Delhi, India. She is B.E (Electonics and Communication-83, Delhi Univ.), M.Tech (Integrated Electronics and Circuits'92, IIT-Delhi), Ph.D (Delhi University). Her research interests include Design space exploration of hardware software architectures, Information Retrieval and Computer Security.

Arabian Journal for Science and Engineering
https://doi.org/10.1007/s13369-019-03992-7

RESEARCH ARTICLE - SPECIAL ISSUE - INTELLIGENT COMPUTING AND INTERDISCIPLINARY APPLICATIONS

A Hybrid-Based Verifiable Secret Sharing Scheme Using Chinese Remainder Theorem

Om Prakash Verma[1] · Nitin Jain[2] · S. K. Pal[3]

Received: 12 December 2018 / Accepted: 19 June 2019
© King Fahd University of Petroleum & Minerals 2019

Abstract
It is not always in the best interests to rely on an individual to have control of entire sensitive information. This has led to the need for secret sharing schemes, which divide secret (key) among many participants or shareholders. To avoid any cheating by any of the shareholders, the need for verifiable secret sharing (VSS) has emerged. In this context, a hybrid approach for VSS scheme is suggested in this paper. The proposed algorithm shares multiple secrets among shareholders, where shareholders are also divided/classified into different levels. Hence, it includes multiple as well as multilevel secret sharing. Secrets can be recovered at intra- or inter-level, where shareholders of higher level can contribute their shares to lower levels. To reduce the complexity, the one-way hash function is used instead of the hard number-theoretic problems. The proposed scheme stands against the dishonest dealer and shareholders. To rule out a typical dishonest strategy of leaking secret information in the valid shares, the concept of dealer leakage resilience is used by reducing the dealer's powers of selecting random values on his own. The execution is also done using cryptographic libraries. Finally, it is demonstrated that the scheme satisfies the security requirements of VSS.

Keywords Verifiable secret sharing schemes · Hash functions · Dealer leakage resilience · Secret recovery · Chinese remainder theorem · Dishonest participants

1 Introduction

The concept of secret sharing (SS) schemes was coined by Shamir [1] and Blakley [2] in 1979. Since then, it has attracted the interest of several researchers. SS has been found valuable in several applications such as witness encryption [3], secure communication [4], and access control [5]. In SS schemes, there are two significant role players, one is the dealer and another is the group of shareholders (participants). The dealer splits the secret into n parts and distributes these shares among n shareholders. These shareholders when combining their shares can recover the secret. The scheme is referred to as a thresholding scheme if the secret can be recovered by combining t out of n ($t \leq n$) shares. However, less than t parts must not reveal any information about the secret. There are some drawbacks in the SS schemes presented in [1, 2] which may act as a constraint for practical usage:

- Fake shares may be distributed by the malicious dealer, and in turn, secret reconstruction is not possible.
- A deceitful shareholder may submit a fake/invalid share, which leads to incorrect share reconstruction, and the true secret would only be known to the deceitful shareholder.
- Need of mutually trusted dealer for the generation and distribution of shares.
- There is a requirement of the private channel for share distribution.

An advancement of SS schemes, known as verifiable secret sharing (VSS) schemes, came into the picture to handle the dishonesty of shareholders mainly or dealer. The dealer may be biased in the distribution of shares or the reconstruction of the secret.

✉ Nitin Jain
garg.nitin007@gmail.com

[1] Department of ECE, Delhi Technological University, Delhi, India

[2] Department of Information Technology, Delhi Technological University, Delhi, India

[3] Directorate of Information Technology & Cyber Security, DRDO, Delhi, India

Published online: 25 July 2019

In traditional SS schemes, it is assumed that the shareholders and the dealer are honest and reliable enough. Though in practical scenarios, the dealer does not trust the players completely, and consequently, it is reasonable to expect that players may not trust the dealer as well. To make SS verifiable, some auxiliary information is to be added that helps the shareholders to verify their respective shares. The shares are that shareholders do not accept the shares if they find them inconsistent or invalid. With the help of VSS schemes, it is possible for the shareholders to verify their shares without having access to the secret and even without revealing their shares. Other flavours of SS schemes include multiple [6], multilevel [7], weighted [8], and protected SS (PSS) schemes [9].

2 Related Work

In multiple SS schemes, there are multiple (say p) secrets instead of a single secret as in traditional SS schemes. To share p secrets, one approach is to run p instances of the simple scheme. However, this seems to be a very naïve way and not a desirable solution due to high computational complexity. So a scheme is desirable if single run [6] can share all p secrets. Recent work in this direction is done by Amroudi et al. [10], where authors obviate the need of a secure channel by encrypting the shares with the NTRU cryptosystem which is a lattice-based and reasonably fast approach. A multivariate polynomial's coefficient is used to share the multi-secret with the verification of shares performed by using the hash function. Another work in this league is a scheme by Meng et al. [11] that uses cellular automata and the hash function for verifiability. Trust management without the dealer is achieved with the help of linear computations with the simultaneous use of parallel computation for efficiency. Another multiple SS scheme is proposed by Tentu et al. [12], where multiple secrets are distributed using discrete logarithm and quadratic residue problem. This scheme is used for the level-ordered access structure. Cheng et al. [13] proposed a verifiable multi-secret sharing based upon the Lagrange polynomial and the public key cryptosystem. They used a linear feedback shift register (LFSR)-based cryptosystem to enhance the efficiency of the scheme. The scheme provides reasonably good security with added efficiency. In recent past, Giri et al. [14] proposed a multi-scheme whose assumption is based upon the geometry in the finite field. The scheme is claimed to be secured as the shares which are shared with the participant, are not the actual values but the shadow values. Liu et al. [15] proposed a multi-secret scheme which proved the failure of asynchronous reconstruction of share given by the Harn and Hsu [16]. They also proved that by getting the reconstruction of any of the secret, rest of the secrets could be obtained illegitimately. They improved the abnormality of the scheme by taking the common pairwise key for a pair of shareholders.

In multilevel or hierarchical SS schemes [7], participants are divided into m different levels and a threshold is associated with each level. For secret recovery, a participant from the targeted level or higher level can contribute in the secret reconstruction. Zhong et al. [17] extended the idea of giving a shadow number to images. Shadow image as a share stops the cheating in the shareholder before the actual image is recovered. They extended the idea of a weighted scheme by giving the higher priority to the shareholder at a higher level; i.e. capabilities of shareholders at different levels are different.

In weighted SS schemes [8], a weight with a positive value is assigned to each participant. Secret reconstruction is possible only when the sum of weights of the authorised subset is equal to or greater than the threshold. In previously described schemes [1–7], each shareholder has unity weight. However, in this, different shareholders may be assigned different weights. The concept of such schemes can be applied directly when there is a need to give more rights to higher rank officials.

In SS schemes, traditionally, to avoid the chances of recovery of a secret by non-shareholders, secure pairwise channels are established among the shareholders via means of a shared key, which are used to exchange the shares. To reduce this computational inefficiency, Harn et al. [9] coined PSS scheme. In this, in addition to the secret reconstruction, the shared key is also established with the help of shares possessed by the shareholders in a pairwise manner. Though this scheme is computationally less efficient than the Shamir's SS scheme, it can be used even if the adversary has unlimited computational power.

Another class of VSS schemes is known as publically VSS (PVSS) scheme [18, 19]. Such VSS schemes possess a unique property that anyone can verify that distributed shares are valid or not, i.e. maliciousness of the shareholders can be handled by this type of scheme. Shareholders receive a valid share but do not submit a valid one during reconstruction. A remarkable PVSS scheme was presented by Behnad et al. [20], where members of the participant can be proven by themselves, avoiding illegal member's participation.

With the proliferation of big data and cloud computing technology and its associated requirement, homomorphic secret sharing schemes were proposed in the recent past. Though the concept of cryptographic homomorphism is ancient, it has been touched by various researchers from time to time. Li et al. in 2018 discussed the various cryptographic primitives which can be used for privacy preservation requirement of various online applications [21]. A scheme by Rajabi et al. [22], whose security is based upon the approximate shortest polynomial problem, exploits homomorphic as well as collision resistance property by taking appropriate Knapsack function. The verification of shares can also

be done using public channels. Another variant of VSS is asynchronous verifiable SS (AVSS) scheme, where fault tolerance in multiparty computation can be handled. Basu et al. [23] proposed an optimistic AVSS scheme, where the payoff cost of failure possesses linearity, i.e. proportional to the number of failures. A different approach which adds non-malleability to SS scheme was proposed by Goyal et al. [24]. With this scheme, if the shares are tempered, then either the original secret can be recovered, or the recovered secret is unrelated to the original secret.

By getting motivation from these requirements, our work proposes a hybrid-based VSS scheme using Chinese remainder theorem (CRT) that is based on hash functions [7, 25, 26]. The proposed scheme also stands against the dishonest dealer and shareholders. To rule out a typical dishonest strategy of leaking secret information in the valid shares, the concept of dealer leakage resilience is used by reducing the dealer's powers of selecting random values on his own. In the results and analysis section, it is confirmed that the scheme adheres to the security requirement of VSS.

The rest of the paper is organised as follows: In Sect. 3, the preliminary background and applications of SS schemes are explained. Section 4 explains the proposed work. Next, Sect. 5 demonstrates the experimental results and also explains how the proposed scheme can be used for Defence application also. Further, Sect. 6 analyses the scheme against various security parameters, where the comparison is also made with the existing schemes. Finally, Sect. 7 concludes the paper.

3 Preliminary Background and Applications of SS Schemes

3.1 Background of SS Schemes

Simple SS schemes are not of much interest in practical scenarios. A threshold value plays an important role. Very first idea in this league is a scheme due to Adi Shamir [1] in 1979. In every SS scheme, there are two phases, one is share generation and another is secret reconstruction. Shamir's scheme is based on Lagrange's polynomial interpolation which satisfies the basic requirements of SS schemes. Shareholders can unlock the secret if t (out of n) or more shares are known. Shamir's scheme is divided into two algorithms, namely share generation and share reconstruction.

3.1.1 Share Generation

In this, the dealer selects a polynomial $f(x)$ (given by (1)) of degree $t-1$ whose coefficients are randomly chosen from a finite field by the dealer,

$$f(x) = a_0 + a_1 * x + a_2 x^2 + \cdots a_{t-1} x^{t-1} \tag{1}$$

Dealer computes a set of n shares $\{f(1), f(2), \ldots, f(n)\}$ and distributes them among the participants through the private channels.

3.1.2 Secret Reconstruction

The secret reconstruction is not done until t parties are involved. For example, as a minimum two points are required to construct the equation of a line, three points are required for formulating a quadratic equation and similarly, t shares are combined to reconstruct equation of degree $t-1$. According to Shamir's scheme, the polynomial reconstruction is done using the Lagrange's polynomial interpolation, i.e.

$$f(x) = \sum_{i=1}^{t} f(i) \prod_{j=1, j\neq i}^{t} (x-j)/(i-j) \tag{2}$$

Another landmark work in this direction was presented by Blakey [2] in the same year where the scheme was based on hyperplane geometry. It can be summarised that as the secret is a specific point in space, each share corresponds to hyperplane and the number of planes intersecting (if more significant than the threshold) reveals the secret.

The notion of verifiability in SS schemes was first presented by Chor et al. [27], where verification of received share is done without any information about the secret. VSS schemes can be interactive and non-interactive [28], where non-interactive schemes are more efficient in comparison with the interactive ones. Initially, interactive schemes were presented in which the dealer and players communicate with each other to check the validity of the shares. This sometimes increases the overhead of the dealer as he has to communicate with N players. Later, non-interactive schemes were introduced which reduced the dealer's overhead (communication). Max Mignotte [29] came with his seminal work in 1983 that was based on CRT and used a particular sequence of integers rather than using an interpolation polynomial for secret construction. Another popular construction is due to Feldman [30] which is verifiable and non-interactive one based on Shamir's scheme. The security is based on discrete logarithm problem (DLP) which is assumed to be computationally secure. To make the scheme unconditionally secure, Pedersen [31] used a commitment to function in his scheme.

On the other hand, if the dealer can get commitment values and break DLP, he can distribute fake shares. In 2008, Kaya et al. [32] proposed another VSS scheme based on CRT and proved its security. They also proposed a joint random secret sharing (JRSS) and proactive SS scheme protocol. In 2010, Harn and Lin [33] defined (n, t, n) the SS scheme based on Pedersen's schemes and presented the notion of strong

VSS and t consistency. They also presented a robust (n, t, n) scheme based on Benaloh [34] scheme. Subsequently, in 2012 to 2014 Meng et al. [35] and Mahmoud [36, 37] proposed different VSS schemes. In the direction of PVSS, different schemes [18–20, 38–45] have been proposed from time to time with different capabilities.

3.2 Applications

There are various applications [3–5, 46] of SS schemes ranging from traditional to contemporary. SS schemes can be used for hierarchal organisations to share a single secret. The proposed scheme (Sect. 4) can be used to share multiple secrets in a multilevel environment with fulfilling the necessary security requirements. Other applications of SS are as follows:

Securing Cryptographic Keys play an essential role in any cryptosystem. In such cases, the key is split into different parts. Each part is termed as a share of the key, and these shares are distributed to all the participants who pool their shares for key construction.

Electronic Voting also called as E-Voting uses electronic systems for casting and counting votes. To avoid plausible dishonesty, SS schemes can be adopted in the E-Voting system. Each vote can be treated as a secret, and shares of the vote are distributed among the authorities who are counting the votes. Now only t authority can access the vote, and it cannot be manipulated by any $t-1$ authorities. SS schemes add security and reliability to the E-Voting system. Another possible application of the electronic system is *E-Auction.* In this system, participants put an offer for the items and allocation is done based on their offered prices.

Similarly, SS can be used for *Threshold Schemes for Multiple Servers as well.* Shares are spread across multiple servers, and even $t-1$ shares do not give any information. The scheme works even if one or two servers meet any failure and the secret can still be recovered.

Distributed Signatures is a mathematical way to authenticate a message. It is generally a hash code of the message, encrypted with a secret key. Sender puts his signature to authenticate the message. If there are multiple co-signers, each of them signs the message one by one according to the priority. However, this is not an efficient way because any co-signer can repudiate. The SS schemes can be adopted in such a scenario. Signing key acts as secret which is shared among all the co-signers. Each share is given to each co-signer, and no one is having complete control over the secret. Minimum t co-signers need to pool their shares for signing key construction. Thus, the scheme is secure and repudiation is not possible.

In the next section, a hybrid-based VSS scheme is proposed and in the subsequent sections, the results are analysed and compared with the existing VSS schemes.

4 Proposed Algorithm

The proposed scheme works for multiple secrets in the multilevel structured environment (hierarchal organisation). In this, the shareholders are divided into z levels $(L_1, L_2, \ldots, L_Z)$ with L_1 and L_Z as the highest and lowest levels, respectively. Each ith level is assumed to have N_i shareholders. For example, if $N_3 = 4$ it implies that there are 4 shareholders at level 3. There is a dealer D who wants to share k secrets $(M_0, M_1, \ldots, M_{k-1})$ among the shareholders, and let t be the threshold of the protocol. Whole protocol is divided into 2 phases: share generation and secret reconstruction. The essential conditions necessary for successful secret reconstruction are:

- The secret can be reconstructed if there are t or more valid shares available.
- The secret cannot be reconstructed if the number of shares is less than t.

Each shareholder keeps $k+t$ values as their shares which are used to reconstruct k secrets. The whole algorithm is explained as below:

4.1 Share Generation

Assume there are k secrets and all are from Z_{p}^* where p is a big prime.

Case 1 Intra-Level Secret Sharing

- D forms a polynomial $f(x)$ of a degree $(t+k-1)$ from Z_{p}^*, i.e.

$$f(x) = \sum_{i=0}^{t+k-1} ai * x^i \bmod p \tag{3}$$

where $a_0 = M_0, a_1 = M_1, \ldots, a_{k-1} = M_{k-1}$ and $a_k, a_{k+1}, \ldots, a_{k+t-1}$ are the private values given by the shareholders to the dealer through a private channel.
- D selects an integer I_0. For each level, a sequence of pairwise co-prime positive integers is selected and made public. Integers at each level equal to the number of shareholders at that level, i.e. $\left(I_1^i, I_2^i, \ldots, I_{N_i}^i\right)$ with $\left(I_1^i < I_2^i < \cdots < I_{N_i}^i\right)$ where $i = 1, 2, \ldots, z$ and greatest common divisor (GCD) of I_0 with every other selected integer should be 1.

- D creates $t+k$ shares of the polynomial, and for each share $f(r)$, dealer forms $f(r)+\delta^i_{x,N_i} * I_0$, where δ^i_{x,N_i} is a random value selected by the dealer for share number x of shareholder N_i at ith level with x varying from 1 to $t+k$. In N_i, N is the number of shareholders at ith level. The value δ^i_{x,N_i} is different for each level and each share. $\left(f(r)+\delta^i_{x,N_i} * I_0\right)$ should lie between

$$\left(I^i_{N_{i-t+2}} * I^i_{N_{i-t+3}} * \cdots * I^i_{N_i}\right) < \left(f(r)+\delta^i_{x,N_i} * I_0\right) < \left(I^i_1 * I^i_2 \cdots I^i_t\right) \quad (4)$$

This is the threshold range for every level, and secrets should lie in this range; otherwise, the algorithm would be inconsistent, i.e. reconstruction can be possible by combing less than t shares. The value to be shared is S^i_{x,N_i}: (S^i_{x,N_i} corresponds to share a number x of the shareholder N_i at ith level with x varying from 1 to $t+k$).

$$S^i_{x,N_i} = \left(f(r)+\delta^i_{x,N_i} * I_0\right) \bmod I^i_{N_i} \quad (5)$$

- Before distributing S^i_{x,N_i}, D computes its hash values and these values are made public so that everyone can access it. Shareholders accept the share if and only if its hash value matches with the previous hash value published by the dealer otherwise discard it. This mechanism checks the dishonesty of the dealer and makes the scheme verifiable. Thus, the dealer is not able to distribute invalid shares.
- The dealer distributes shares S^i_{x,N_i}. Similarly, $t+k$ polynomial values are shared among all the shareholders at each level.

Case 2 Inter-Level Secret Sharing

For inter-level SS, D, needs to select another parameter $I^i_{N_{i,j}}$, where the shareholder N_i contributes his share to jth level for secret reconstruction with.

$$I^j_t < I^i_{N_{i.j}} < I^j_{N_{j-t+2}} \quad (6)$$

Then, the dealer computes $\Delta S^i_{x,N_{i,j}}$

$$\Delta S^i_{x,N_{i,j}} = f(r)+\delta^i_{x,N_i} * I_0 - S^i_{x,N_i} \quad (7)$$

with a share of the shareholder in inter-level sharing as $S^i_{x,N_i} + \Delta S^i_{x,N_{i,j}}$.

4.2 Secret Reconstruction

A system of equations is formed based on the distributed shares. Dealer D accepts shares only if the share is valid, which is verified using the hash value published by the D before. An equation which is formed which is given as:

Case 1 Intra-Level Secret Sharing

$$\delta^i_{x,N_i} * I_0 \bmod I^i_{N_i} \quad (8)$$

Case 2 Inter-Level Secret Sharing

$$S^i_{x,N_i} + \Delta S^i_{x,N_{i,j}} \bmod I^i_{N_i,N_j} \quad (9)$$

Using CRT, a unique solution for $X = f(r)+\delta x, i * I_0$, $f(r)$ can be reconstructed by

$$f(r) = x \bmod I_0 \quad (10)$$

After getting all the polynomial shares by CRT, the following equation is used to reconstruct the polynomial

$$f(x) = \sum_{i=1}^{t} f(i) \prod_{j=1, j\neq i}^{t} (x-j)\big/(i-j) \bmod p = a_0 + a_1 \cdot x^1 + \cdots a_{k+t-1} \cdot x^{k+t-1} \quad (11)$$

Thus, the authorised set of shareholders reconstructs the k secrets.

The proposed scheme is demonstrated in Fig. 1 which shows 3 levels with 3 shareholders at each level.

5 Results and Discussion

5.1 Experimental Results

The above proposed scheme described in Sect. 4 is implemented in C/C++ using GMP (GNU Multiple Precision) and NTL (Number Theory Library) libraries and tested on 3-GHz third-generation system. GMP is a free library which is multi-precision and can used for various types of operations on signed integers, floating point numbers, and rational numbers. The richness of function, friendly interface, and freely availability makes it so popular and useful. The limit of precision just depends upon machine not on the library. The application includes cryptography and its application, security over the internet, algebraic number theory, and many more. For better insight, implementation results are presented for small numbers and the algorithm is tested for large numbers as well. For demonstration, shareholders are divided into 3 levels ($z=3$), namely L_1, L_2, and L_3 with 3, 4, and 7 as the number of shareholders at respective levels, i.e. $N_1=3$, $N_2=4$, $N_3=7$. The threshold (t) and the prime (p) being considered are 3 and 563, respectively. Number of secrets

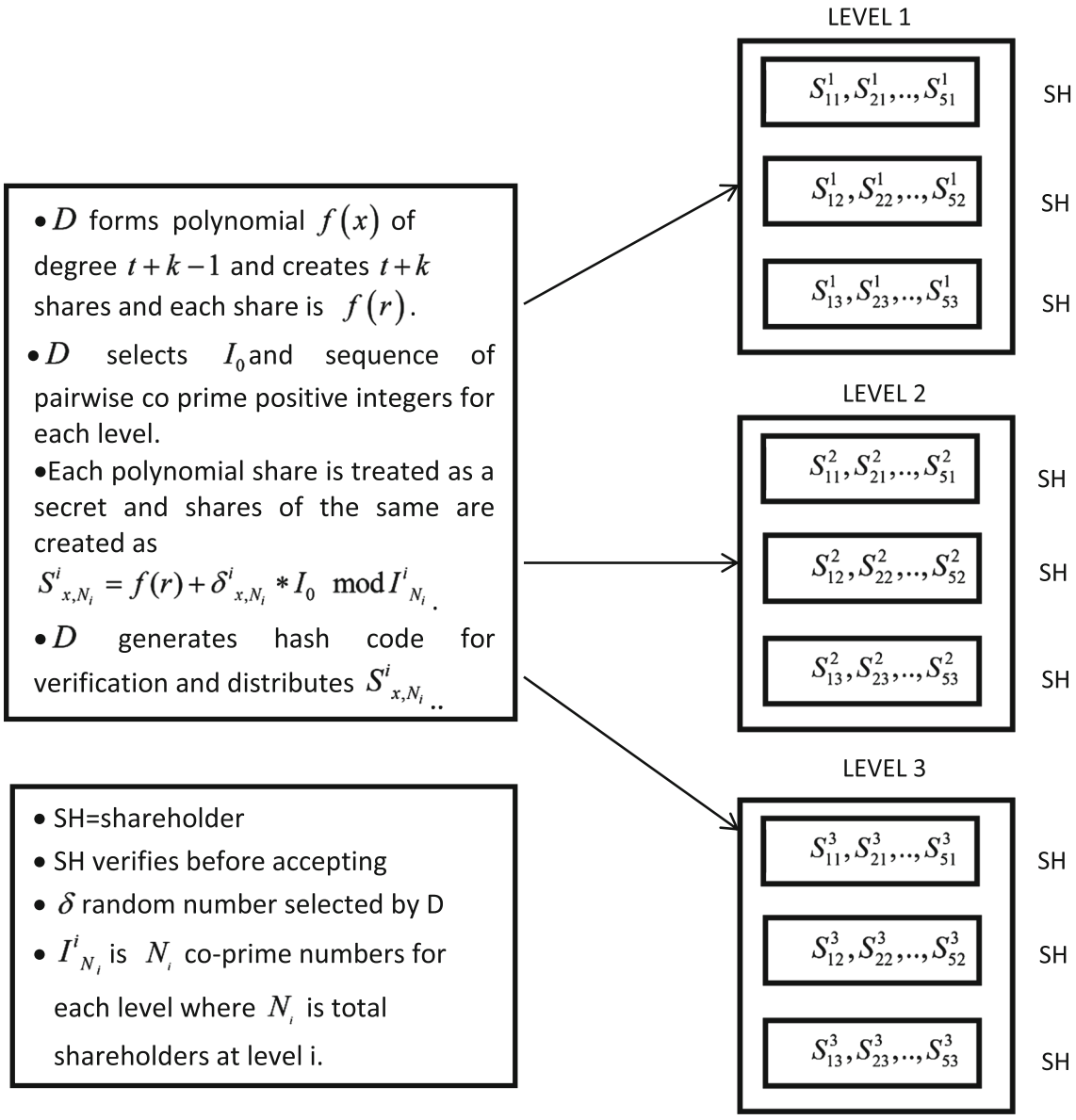


Fig. 1 Proposed algorithm

to be shared is 2 ($k = 2$), where $K_0 = 3$ and $K_1 = 2$ and the coefficient values provided by an authorised set of players are 2, 1, 0 (authorised set used for secret reconstruction comprises 3 players).

(1) Dealer forms the polynomial of a degree $t + k - 1$ using the above values, i.e.

$$f(x) = 3 + 2 * x + 2 * x^2 + 1 * x^3 + 0 * x^4 \quad (12)$$

(2) Dealer selects $I_0 = 863$ and the sequence of pairwise co-prime integers selected for each level are

- For level 1: $I_1 = 137$, $I_2 = 139$, $I_3 = 250$, and threshold range for this level is (34750, 4760750)
- For level 2: $I_1 = 293$, $I_2 = 307$, $I_3 = 313$, $I_4 = 319$, and threshold range for this level is (99847, 28154663)
- For level 3: $I_1 = 229$, $I_2 = 233$, $I_3 = 239$, $I_4 = 241$, $I_5 = 277$, $I_6 = 281$, $I_7 = 283$, and threshold range for this level is (79523, 12752323)

(3) Dealer creates $t + k$ shares: $f(1) = 8$, $f(2) = 23$, $f(3) = 54$, $f(4) = 107$, $f(5) = 188$, and selects $\delta_{x,i}$ as a value for each level which is shown in Table 1.

Therefore, all 5 shares of a 1st shareholder are:

$8 + 550 * 863 \bmod 137 = 90$

$23 + 558 * 863 \bmod 137 = 22$

$54 + 510 * 863 \bmod 137 = 3$

Table 1 $\delta_{x,i}$ values selected by the dealer

Levels	$\delta_{1,i}$	$\delta_{2,i}$	$\delta_{3,i}$	$\delta_{4,i}$	$\delta_{5,i}$
Level 1	550	558	510	620	456
Level 2	9864	8824	8999	9345	9500
Level 3	10,946	10,567	11,001	10,765	10,899

$$\begin{aligned} &107 + 620 * 863 \bmod 137 = 45 \\ &188 + 456 * 863 \bmod 137 = 115 \end{aligned} \tag{13}$$

Similarly, shares of other shareholders are calculated which are shown in Table 2.

Case 1 Intra-Level Secret Sharing

While recovering the secret from level 2, 3 out of 4 shareholders need to contribute their shares. Say, first 3 are taking part in the protocol. Following the system of equations needs to be solved using CRT:

$$\begin{aligned} &X = 111 \bmod 293 \\ &X = 144 \bmod 307 \\ &X = 292 \bmod 313 \end{aligned} \tag{14}$$

This gives $X = 8$. In the same way, equations can be formed with other shares of the shareholders and results can be obtained accordingly.

Case 2 Inter-Level Secret Sharing

Considering secret reconstruction done at level 3, a 1st shareholder of each level is contributing their shares for reconstruction. Dealer selects two values (because 2 out of 3 shares belong to other levels) between 241 and 277 which are co-prime to one another. Say the values are 253 and 263; following system of equations is formed for secret recovery:

$$\begin{aligned} &X = 90 + 55 \bmod 253 \\ &X = 111 + 124 \bmod 263 \\ &X = 156 \bmod 229 \end{aligned} \tag{15}$$

Solving these equations using CRT, we get $X = 8$ and similarly other shares are obtained. Further, these values are used in Lagrange's interpolation to reconstruct the polynomial

$$f(x) = 3 + 2 * x + 2 * x^2 + 1 * x^3 + 0 * x^4 \tag{16}$$

Moreover, the secrets are recovered.

5.2 Application of the Proposed Scheme

The proposed scheme can be advantageous to share multiple secrets in a multilevel environment (for organisations having hierarchal structures). Considering an example of Indian Army, suppose a Colonel is having some secrets (secret keys/passwords) and he is on leave or some special mission for some days. Practically, it is not advisable to hand over the secrets to a single officer (superior or subordinate). So, he would make shares of the secrets and hand it over to officers of various ranks. He may give some shares (these shares include shares formed by splitting multiple secrets to be circulated) to higher rank officers (Lieutenant Colonel or Brigadier) and others to peers or subordinates (Major, Captain and Lieutenant). Now, if an emergency arises for secret reconstruction at Captain level, then, the officer only at a peer or higher rank can contribute in share reconstruction. The secret is reconstructed, provided the threshold condition is satisfied. This algorithm can be used in the case when a higher rank officer (say Colonel) does not want entities or members at lower levels (Major, Captain, and Lieutenant) to recover secrets on their own without any member of lower levels. For this, he can set the threshold value more than the number of entities present at that (lower) level or another alternative is to provide more shares to entities at a higher level and less number of shares to entities at lower levels.

6 Security Analysis and Comparison

The proposed work is analysed in this section, and a comparison with some existing schemes is also performed.

6.1 Security Analysis

Traceability Algorithm is said to be traceable when it is possible to find out whether any participant during the reconstruction phase has submitted any invalid or fake share or not.

Proof Let $f(i)$ be the original valid share and $f'(i)$ is the fake or invalid share. If any participant sends $f'(i)$ to the dealer instead of $f(i)$, then the dealer does not accept the share because

$$H(f(i)) \neq H\left(f'(i)\right) \tag{17}$$

Here, H is one-way hash function and it is complicated to find 2 values that result in the same hash value. Thus, the algorithm is traceable.

Robustness Scheme is said to be robust if all the secrets can be recovered by pooling t or more shares. Use of Lagrange Interpolation has made the scheme more robust. Any t honest players can unlock the shared secret.

Table 2 Shares of shareholders

Shareholders	1st share	2nd share	3rd share	4th share	5th share
1st shareholder of level 1	90	22	3	45	115
2nd shareholder of level 1	112	81	110	17	68
3rd shareholder of level 1	158	77	184	167	216
1st shareholder of level 2	111	65	226	17	255
2nd shareholder of level 2	144	0	12	259	253
3rd shareholder of level 2	292	158	35	84	279
4th shareholder of level 2	125	286	136	203	69
1st shareholder of level 3	156	106	35	1	79
2nd shareholder of level 3	120	190	99	126	48
3rd shareholder of level 3	170	60	120	133	180
4th shareholder of level 3	170	145	204	234	36
5th shareholder of level 3	152	227	19	276	213
6th shareholder of level 3	29	51	51	161	112
7th shareholder of level 3	149	235	116	261	237

Confidentiality Scheme holds confidentiality if even $t-1$ players are not able to reveal the secret. Assume $t-1$ participants are available for secret recovery and product of their moduli is X'. These $t-1$ shareholders use CRT to recover a secret. Suppose they obtained a value S'. The relation between the original secret and the recovered secret is

$$S = S' + \delta * X' \tag{18}$$

Here, S is the original secret. Predicting the correct value of δ to reach the original secret is very difficult. Thus, even with $t-1$ shares, the scheme does not leak any information about the secret.

Consistency Algorithm holds consistency if any set of valid shares of the secret reveals the same secret. Here in the proposed algorithm, consistency is achieved due to Lagrange's interpolation and whether the share is valid or not is verified through one-way hash function.

Dealer Leakage Resilient (DLR) Dealer is said to be dishonest if he subliminally leaks the information in the valid shares. This dishonest strategy allows the dealer to preserve consistency in the system and helps the attacker to unlock secret before reconstruction phase from the leaked information. The system exhibits DLR-VSS property if the attacker does not gain information about the secret before the reconstruction phase.

Proof The DLR-VSS property is achieved by taking the power of randomness from the dealer. The dealer does not have the capability of employing randomness in the system. By this, the dealer will no longer be able to hide information because no value is selected by his own choice.

Salted Hashing can be used in place of simple hashing. In salted hashing, a random number, referred as salt, is added to the share before using one-way hash functions. Salted hashing ensures that no two similar secrets yield similar hash codes. However, the only dealer can verify shares submitted by shareholders. It is assumed that the dealer is honest and he is not distributing invalid shares. When we randomise the hashes, rainbow tables, lookup tables, and reverse lookup tables are no more an effective tool. For the pre-computation of rainbow or lookup table, salt needs to be known in advance and this is not possible.

Another possible method to use salted hashing and still verification is possible from both ends, i.e. shareholders can verify shares before accepting it from dealer and dealer also can verify share before accepting it from shareholders before reconstruction phase. This can be achieved by treating salt (or random number) as one of the secrets. The constant term of the polynomial will be the salt, and degree of the polynomial is $t+k$.

Knowledge of Number of Shares If t and k are not publicly known values, then, it is desirable that adversary must not get any information about some secrets (in case of multi-secret schemes) just by looking at the number of shares of each shareholder. This property is achieved by distributing $t+k$ shares instead of k shares, and t and k are kept a secret, so adversary is not able to access these values. Table 3 shows the comparison of the schemes [7, 25, 26, 47–50] by security assumptions. The acronyms R, C, V, and T stand for robustness, confidentiality, consistency, and traceability, respectively.

Table 3 shows that the proposed approach satisfies all the properties (R, C, V, and T) of VSS schemes. Therefore, it is confirmed that the scheme is verifiable.

In Table 4, the proposed algorithm is compared with other schemes [7, 25, 26, 47–50] by communication cost over secure and insecure channels. Communication cost over

Table 3 Comparisons through security property

Scheme No.	Robustness (R)	Confidentiality (C)	Consistency (V)	Traceability (T)
[7]	Yes	Yes	No	Yes
[47]	Yes	Yes	No	Yes
[48]	Yes	Yes	No	Yes
[25]	Yes	Yes	Yes	Yes
[49]	Yes	Yes	No	Yes
[26]	Yes	Yes	No	Yes
[50]	Yes	Yes	Yes	Yes
Proposed	Yes	Yes	Yes	Yes

Table 4 Comparison through communication cost

Scheme	Distribution cost over secure channels	Reconstruction cost over secure channels	Communication cost over insecure channels	Security assumption
[7]	$1024\,(n_1 + n_2 + \cdots n_z)$	$1024 * t$	$1024\,(n_1 + n_2 + \cdots n_z)$	Unconditionally secure
[47]	–	$160\,t$ or $160\,k$	$1184n + 160n + 1024t$ or $1184n + 160n + 1024k$	DLP
[48]	$1184\,n$	$160\,t$ or $160\,k$	$1024n + 160n$	RSA and DLP
[25]	$320\,n$	$160\,t$ or $160\,k$	$C * n$	Hashing
[49]	–	$160\,t$ or $160\,k$	$1184n + 1184n + 160n$ or $1184n + 1184n + 160k$	RSA and DLP
[26]	$160\,n$	$160\,k * t$	$160 * n * k + C * n * k$	Hashing
[50]	$320\,n$	$2048t$	$2048\,(n + 1) + 2048n$	DLP
Proposed	$160 * (t + k)*$ $(n_1 + n_2 + \cdots n_z)$	$160\,k * t$	$(160 + C)*$ $(n_1 + n_2 + \cdots n_z)$	Hashing

secure channels is analysed separately for both share distribution and share reconstruction. Here, n, t, and C, respectively, denote the number of shareholders, threshold, and any constant number. Communication cost over a secure and insecure channel is calculated using p (1024 bits), q(1024 bits), and $N = p * q$ (1184 bits). In addition to communication cost, security assumption of different schemes is also analysed.

In [7], variables n_1, n_2, and n_3 used are some shareholders at different levels, where z is the total number of levels. This scheme is multilevel secret sharing and uses Chinese remainder theorem. Consider there are z levels. Here, dealer publishes sequences of co-prime numbers equal to the number of shareholders for each level which are of 1024 bits each. This adds to the communication cost of 1024 $(n_1 + n_2 + \cdots n_z)$ bits over insecure channels. Dealer computes shares for n shareholder and distributes them which are of 1024 bit each which makes the communication cost over the secure channel to $1024 * n$ bits. Then, t shareholders collaborate to reconstruct the secret. Thus, the communication cost for reconstruction is $1024 * t$ bits.

In [47], each participant selects his secret shadow of 1184 bits and sends it to the dealer through secure channel which makes the communication cost over the secure channel to $1184\,n$ bits. There are two possible cases, first is when $t > k$, in this case, dealer forms polynomial of degree $(t - 1)$. He creates shares of the polynomial of 160 bits each and publishes them $(160 * n)$. Dealer computes t values of 1024 bits that are used in verification phase and publishes them $(1024 * t)$. t shares of 160 bit each are pooled to recover the secret which makes cost of reconstruction as $160 * t$. Second case arises, when $t < k$. This time dealer forms polynomial of degree $(k - 1)$. He creates shares of the polynomial of 160 bits each and publishes them $(160 * n)$. Dealer computes t value of 1024 bits that are used in verification phase and publishes them $(1024 * t)$. k shares of 160 bit each are pooled to recover the secret which makes cost of reconstruction to $160 * k$ in the second scenario.

In [48], which is multi-secret sharing scheme, where each participant selects his secret shadow of 1184 bits and sends it to the dealer through secure channel which makes the communication cost over the secure channel to $1184 * n$ bits. After some computation, dealer publishes a value of 1024 bits for each participant, which adds $1024 * n$ bits to the communication cost over insecure channels (public channel); then,

Table 5 Comparison through various parameters

Property	[7]	[25]	[50]	[26]	[51]	Proposed algorithm
Dealer publishes the share	No	No	Yes	Yes	No	No
Hash for verifiability	No	Yes	No	Yes	–	Yes
Modular exponentiation or DLP	No	No	Yes	No	No	No
Multi-use scheme	–	No	No	Yes	No	Yes
Can verify dealer's honesty	No	Yes	Yes	No	No	Yes
Can verify shareholder's honesty	No	Yes	Yes	Yes	No	Yes
Has unconditional security	Yes	No	No	No	Yes	No
Outside adversary does not know number of secrets	–	Yes	–	No	No	Yes
Secret revealing order	–	All at a time	All at a time	Any	Any	Any
Conspiracy attack resistance	Yes	Yes	Yes	Yes	Yes	Yes
SETUP attack resistance	No	No	Yes	No	No	Yes

dealer uses public channel to distribute shares of 160 bits for each shareholder $(160 * n)$. t or k shareholders submit their shares for secret reconstruction. Thus, the reconstruction cost becomes $160 * t$ if secrets are less than threshold and it is $160 * k$, if secrets are more than threshold.

In [25], there are k secrets to be shared. There are two cases. In the first case, $t > k$, in this dealer forms 2 polynomials of degree $(t - 1)$. One is used to generate shares of multiple secrets, and other is used for verification phase. Dealer computes n shares of 160 bits each, to each polynomial and send it via secure channel to each shareholder which makes the communication cost over secure channel to $(2n*160)$ bits. Then, he publishes (uses insecure channel) the one hash code for the two shares of constant length C which add $C * n$ communication cost over insecure channels. Now, in the reconstruction phase, combiner/dealer combines t out of n shares of 160 bit each of 1st polynomial to recover the secret, that adds communication cost of $160 * t$ 160*t over secure channel. In the second case, $t < k$, here dealer forms polynomials of degree $(k - 1)$. In this case, the communication cost for reconstruction of secrets is $160 * k$.

In [49], it is also a multi-secret sharing scheme, where participants select their secret shadow of 1184 bits and send it to the dealer via insecure channel, which leads to cost of $1184 * n$. Dealer publishes n values of 1184 bits $(1184 * n)$, which are used for verification. Dealer computes shares of 160 bits for the shareholders which are also published which makes the communication cost over in secure channel to $160 * n$ bits. Then, t or k shares are collaborating to reconstruct secrets, which depend on the value of k. Thus, the communication cost in reconstruction phase will be $160 * t$ or $160 * k$ bits over secure channel.

In Scheme [26], it is multi-secret sharing scheme which uses hashing for verification. The dealer sends a private value of 160 bits each, to each shareholder via secure channel which makes the communication cost over secure channel to $160*n$ bits. He publishes k shares of 160 bit each $(160 * n * k)$ and their hash codes of constant length $(C * n * k)$ for each participant. Then, t shares of k secrets are combined for the reconstruction which makes cost of reconstruction to $160 * k * t$.

Here in [50], each of n shareholders sends 2 private values via a secure channel, which are of 160 bit each, to form 2 polynomials makes $2n * 160$ cost over secure channel. Then, dealer generates and publishes (insecure channel) commitment value, each of 1024 bits for the $2n$ values, which leads to $2n * 1024$ cost over insecure channel. Now, dealer forms 2 polynomials, each of degree $t - 1$. After this, dealer computes n shares to from each polynomial and publishes them. This adds $2n * 1024$ bits to communication cost over insecure channel. In reconstruction phase, combiner/dealer combines t out of n shares of 1024 bits each to recover the secret that adds communication cost of $2t * 1024$ over secure channel.

In the proposed algorithm (multilevel and multiple secret SS scheme), z levels are considered with n_i shareholders for each level (i varying from 1 to z). Dealer forms a polynomial of degree $t + k$, and for each level, publishes a sequence of co-prime numbers which are equal to the number of shareholders at that level $((n_1 + n_2 + \cdots + n_z) * 160$ bit). Each shareholder sends a private value to the dealer through a secure channel which adds $(160 * n)$ bits to the communication cost. Dealer forms $t + k$ shares of this polynomial and creates shares of a polynomial using CRT, and these shares which are of 160 bits, are distributed via a secure channel to shareholders $(160 * (t + k) * (n_1 + n_2 + \cdots + n_z))$. He also publishes the hash code $(C * (n_1 + n_2 + \cdots + n_z))$ of all the shares for verification. In the reconstruction phase, $t*k$ shares are used, which makes reconstruction cost $160 * t * k$ bits.

Table 5 shows the comparison of our scheme with other schemes [7, 25, 26, 50, 51] concerning various parameters mentioned in the table. Some of the properties are explained below:

- The scheme is multi-use if shares of participants are different for different secrets.
- The algorithm can resist conspiracy attack if $t-1$ corrupt shareholders cannot unlock the secret. A conspiracy-resistant scheme ensures that the reconstruction of recovered secret does not give information about open secrets.
- Secretly Embedded Trapdoor with Universal Protection (SETUP) is a technique, where an attacker breaks the security of the system, secret information is leaked, but other parties of the protocol are not able to detect this malicious behaviour. All VSS schemes are not SETUP resilient.
- The scheme is unconditionally secure if its security does not depend on any mathematical construct. It is said to be secure even if the adversary has unbounded computational power.

7 Conclusion and Future Directions

SS is an important sphere of Information Security and is attracting a lot of research interest these days. The primary objective is to develop efficient schemes that are secured and can be deployed practically. In this context, a hybrid-based VSS scheme is proposed in this paper which can be used to share multiple secrets in a multilevel environment. A single run of the scheme can share multiple secrets at different levels. The proposed algorithm holds for all requirements of VSS (Table 5), and the scheme is computationally efficient too (Table 4). The scheme also exhibits the property of dealer leakage resilience which is achieved by restricting the dealer to employ randomness. Consequently, the dealer is not able to hide secret information in the share of the shareholders. So, the proposed scheme to work in an environment where dealers and/or shareholders are not honest and also when they are not mutually trusted. Some promising future work directions are to find a method in which each level or layer can have different thresholds instead of a global threshold and also to find a way to distribute one master share instead of multiple shares for multiple secrets.

References

1. Shamir, A.: How to share a secret. Commun. ACM **22**(11), 612–613 (1979)
2. Blakley, G.R.: Safeguarding cryptographic keys. In: Proceedings of the National Computer Conference, vol. 48, pp. 313–317, June 1979
3. Garg, S.; Gentry, C.; Halevi, S.; Wichs, D.: On the implausibility of differing-inputs obfuscation and extractable witness encryption with auxiliary input. Algorithmica **79**(4), 1353–1373 (2017)
4. Martínez-Peñas, U.: Communication efficient and strongly secure secret sharing schemes based on algebraic geometry codes. arXiv:1610.06082 (2016)
5. Peng, K.: Threshold distributed access control with public verification: a practical application of PVSS. Int. J. Inf. Secur. **11**(1), 23–31 (2012)
6. Binu, V.P.; Sreekumar, A.: Threshold multi secret sharing using elliptic curve and pairing. arXiv:1603.09524 (2016)
7. Harn, L.; Fuyou, M.: Multilevel threshold secret sharing based on the Chinese remainder theorem. Inf. Process. Lett. **114**(9), 504–509 (2014)
8. Iftene, S.; Boureanu, I.C.: Weighted threshold secret sharing based on the Chinese remainder theorem. Sci. Ann. Cuza Univ. **15**(EPFL-ARTICLE-174320), 161–172 (2005)
9. Amroudi, A.N.; Zaghain, A.; Sajadieh, M.: A verifiable (k, n, m)-threshold multi-secret sharing scheme based on NTRU cryptosystem. Wirel. Pers. Commun. **96**(1), 1393–1405 (2017)
10. Harn, L.; Hsu, C.F.; Xia, Z.; Zhou, J.: How to share secret efficiently over networks. Secur. Commun. Netw. **2017**(4), 1–6 (2017)
11. Li, M.; Yu, J.; Hao, R.: A cellular automata based verifiable multi-secret sharing scheme without a trusted dealer. Chin. J. Electron. **26**(2), 313–318 (2017)
12. Tentu, A.N.; Basit, A.; Bhavani, K., Venkaiah, V.C.: Multi-secret sharing scheme for level-ordered access structures. In: International Conference on Number-Theoretic Methods in Cryptology, pp. 267–278. Springer, Cham, September 2017
13. Hu, C.; Liao, X.; Cheng, X.: Verifiable multi-secret sharing based on LFSR sequences. Theoret. Comput. Sci. **445**, 52–62 (2012)
14. Duari, B.; Giri, D.: An ideal and perfect (t, n) Multi-secret sharing scheme based on finite geometry. In: Information Technology and Applied Mathematics, pp. 85–94. Springer, Singapore (2019)
15. Zhang, T.; Ke, X.; Liu, Y.: (t, n) multi-secret sharing scheme extended from Harn-Hsu's scheme. EURASIP J. Wirel. Commun. Netw. **2018**(1), 71 (2018)
16. Harn, L.; Hsu, C.F.: (t, n) multi-secret sharing scheme based on bivariate polynomial. Wirel. Pers. Commun. **95**(2), 1–10 (2017)
17. Liu, Y.N.; Zhong, Q.; Xie, M.; Chen, Z.B.: A novel multiple-level secret image sharing scheme. Multimed. Tools Appl. **77**(5), 6017–6031 (2018)
18. Dehkordi, M.H.; Ghasemi, R.: A lightweight public verifiable multi secret sharing scheme using short integer solution. Wirel. Pers. Commun. **91**(3), 1459–1469 (2016)
19. Mashhadi, S.: Secure publicly verifiable and proactive secret sharing schemes with general access structure. Inf. Sci. **378**, 99–108 (2017)
20. Behnad, A.; Eghlidos, T.: A new, publicly verifiable, secret sharing scheme. Sci. Iran. **15**(2), 246–251 (2008)
21. Li, R.; Xiao, Y.; Zhang, C.; Song, T.; Hu, C.: Cryptographic algorithms for privacy-preserving online applications. Math. Found. Comput. **1**(4), 311–330 (2018)
22. Rajabi, B.; Eslami, Z.: A verifiable threshold secret sharing scheme based on lattices. Inf. Sci. (2018). https://doi.org/10.1016/j.ins.2018.11.004
23. Basu, S.; Tomescu, A.; Reiter, M.; Malkhi, D.: Asynchronous verifiable secret-sharing protocols on a good day. arXiv:1807.03720 (2018)
24. Goyal, V.; Kumar, A.: Non-malleable secret sharing for general access structures. In: Annual International Cryptology Conference, pp. 501–530. Springer, Cham, August 2018
25. Shao, J.: Efficient verifiable multi-secret sharing scheme based on hash function. Inf. Sci. **278**, 104–109 (2014)
26. Das, A.; Adhikari, A.: An efficient multi-use multi-secret sharing scheme based on hash function. Appl. Math. Lett. **23**(9), 993–996 (2010)
27. Chor, B.; Goldwasser, S.; Micali, S.; Awerbuch, B.: Verifiable secret sharing and achieving simultaneity in the presence of faults. In: 26th Annual Symposium on Foundations of Computer Science, pp. 383–395. IEEE, October 1985

28. Bai, G.; Damgård, I.; Orlandi, C.; Xia, Y.: Non-interactive verifiable secret sharing for monotone circuits. In: International Conference on Cryptology in Africa, pp. 225–244. Springer, Cham, April 2016
29. Mignotte, M.: How to share a secret. In: Workshop on Cryptography, pp. 371–375. Springer, Berlin, Heidelberg, March 1982
30. Feldman, P.: A practical scheme for non-interactive verifiable secret sharing. In: 28th Annual Symposium on Foundations of Computer Science, pp. 427–438. IEEE, October 1987
31. Pedersen, T.P.: Non-interactive and information-theoretic secure verifiable secret sharing. In: Annual International Cryptology Conference, pp. 129–140. Springer, Berlin, Heidelberg, August 1991
32. Kaya, K.; Selçuk, A.A.: A verifiable secret sharing scheme based on the Chinese remainder theorem. In: International Conference on Cryptology in India, pp. 414–425. Springer, Berlin, Heidelberg, December 2008
33. Harn, L.; Lin, C.: Strong (n, t, n) verifiable secret sharing scheme. Inf. Sci. **180**(16), 3059–3064 (2010)
34. Benaloh, J.C.: Secret sharing homomorphisms: keeping shares of a secret secret. In: Conference on the Theory and Application of Cryptographic Techniques, pp. 251–260. Springer, Berlin, Heidelberg, August 1986
35. Meng, X.; Li, Y.: A verifiable dynamic threshold key management scheme based on bilinear pairing without a trusted party in mobile ad hoc network. In: 2012 IEEE International Conference on Automation and Logistics (ICAL), pp. 315–320. IEEE, August 2012
36. Al Mahmoud, Q.: Polynomial differential-based strong (n, t, n)-verifiable secret sharing. IET Inf. Secur. **7**(4), 312–317 (2013)
37. Mahmoud, Q.A.: A novel verifiable secret sharing with detection and identification of cheaters. Int. J. Math. Sci. Comput. (IJMSC) **2**(2), 1–13 (2016)
38. Stadler, M.: Publicly verifiable secret sharing. In: International Conference on the Theory and Applications of Cryptographic Techniques, pp. 190–199. Springer, Berlin, Heidelberg, May 1996
39. Fujisaki, E.; Okamoto, T.: A practical and provably secure scheme for publicly verifiable secret sharing and its applications. In: International Conference on the Theory and Applications of Cryptographic Techniques, pp. 32–46. Springer, Berlin, Heidelberg, May 1998
40. Boudot, F.; Traoré, J.: Efficient publicly verifiable secret sharing schemes with fast or delayed recovery. In: International Conference on Information and Communications Security, pp. 87–102. Springer, Berlin, Heidelberg, November 1999
41. Ruiz, A.; Villar, J.L.: Publicly verifiable secret sharing from Paillier's cryptosystem. WEWoRC **74**, 98–108 (2005)
42. Heidarvand, S.; Villar, J.L.: Public verifiability from pairings in secret sharing schemes. In: International Workshop on Selected Areas in Cryptography, pp. 294–308. Springer, Berlin, Heidelberg, August 2008
43. Jhanwar, M.P.: A practical (non-interactive) publicly verifiable secret sharing scheme. In: International Conference on Information Security Practice and Experience, pp. 273–287. Springer, Berlin, Heidelberg, May 2011
44. Wu, T.Y.; Tseng, Y.M.: A pairing-based publicly verifiable secret sharing scheme. J. Syst. Sci. Complex. **24**(1), 186–194 (2011)
45. Shil, A.B.; Blibech, K.; Robbana, R.; Neji, W.: A new PVSS scheme with a simple encryption function. arXiv:1307.8209 (2013)
46. Beimel, A.: Secret-sharing schemes: a survey. In: International Conference on Coding and Cryptology, pp. 11–46. Springer, Berlin, Heidelberg, May 2011
47. Shao, J.; Cao, Z.: A new efficient (t, n) verifiable multi-secret sharing (VMSS) based on YCH scheme. Appl. Math. Comput. **168**(1), 135–140 (2005)
48. Dehkordi, M.H.; Mashhadi, S.: An efficient threshold verifiable multi-secret sharing. Comput. Standards Interfaces **30**(3), 187–190 (2008)
49. Zhao, J.; Zhang, J.; Zhao, R.: A practical verifiable multi-secret sharing scheme. Computer Stand. Interfaces **29**(1), 138–141 (2007)
50. Olimid, R.F.: Dealer-leakage resilient verifiable secret sharing. IACR Cryptol. ePrint Arch. **2014**, 735 (2014)
51. Harn, L.: Secure secret reconstruction and multi-secret sharing schemes with unconditional security. Secur. Commun. Netw. **7**(3), 567–573 (2014)

A Study on Engineering Behaviour of Geogrid Reinforced Clay

Prachi Lingwal
Research Scholar,Delhi Technological University,Delhi, India
prachilingwal57@gmail.com

Shristi Negi
Research Scholar,Uttarakhand Technical University,Dehradun, India
civil.shristinegi@dbgidoon.ac.in

Ankur Mudgal
Research Scholar,Delhi Technological UniversityDelhi, India
ankur.mud@gmail.com

Abstract — Construction work done on cohesive soil have low shear strength and load carrying capacity. Thus, causing reduction in the durability of structure and leading to fail in near future. Earlier to stabilize these types of soil for carrying out construction activity many techniques were used like replacing the top most weak cohesive soil with much stronger granular fill, increasing the footing size, incorporating natural fibers in the soil. However, these conventional methods were not reliable and durable for longer time. Hence use of artificial fibers known as geosynthetic material emerged rapidly. The study shows laboratory results of clay reinforced with geogrid, where load is passed in the soil through a surface square footing having dimensions 75mm×75mm. The parameters analysed in the study includes the vertical spacing of first geogrid layer from footing base (u), optimum depth of effective reinforcement (d), number of geogrid layers (N). Testing was done with geogrid in a steel tank of dimension 750 mm × 450 mm × 600 mm with acrylic sheet along one of the longitudinal sides of tank to get the visual observation of the settlement. The experimental results it is observed that clay soil along with geogrid increases the load carrying capacity of clay when compared to unreinforced clay. Optimum spacing between base of footing and first geogrid is between to 0.34B-0.51B. Depth of effective reinforcement(d) was near by 1.36B with optimum geogrid layers (N) as 4.

Keywords—Cohesive soil; Geogrid; Bearing Capacity; Footing

I. INTRODUCTION

Construction of structure on clay is very difficult due to its low load carrying capacity and settlement over the years. It is low in shear strength and compressible having strength less than 40 kPa and can get easily molded with little pressure. To handle this problem different materials that increase the strength of clay and help it to lift the load coming, are incorporated along. Earlier, the top most clay was removed and was replaced with thick granular full layer or the size of the footing was increased to distribute the load to a larger area thus reducing the pressure. But these techniques could not achieve the actual strength required and moreover were not economical. In 20th century much stable material called polymer was introduced that replaced lesser stable techniques. With time the polymeric material like geosynthetic were used in many engineering works as a reinforcement in mechanically stabilized earth wall, slopes, pavements, earth structure, reinforced foundation.

From the past literatures it is noted that geosynthetic material improves the properties of soil. One of the noticeable works in the area was done by [1] who suggested to use Bearing Capacity Ratio (BCR) technique.

Thereafter the focus area was to work on parameters that would influence the BCR. Reference [9] investigated model strip footing placed over soft clay reinforced with heat-bonded nonwoven polypropylene geotextile. Reference [3] examined footing width effect on sand and concluded that BCR was practically constant when the width was equal to or more than 130 mm to 140 mm. Reference [4] studied improvement of clay subgrade reinforced with geogrid. Optimum distance between footing base and first reinforcement was found at 0.175B. Reference [13] investigated bearing capacity of sand reinforced with geogrid. It was observed that optimum embedment depth is at which bearing capacity is maximum with single reinforcement layer. Reference [12]) observed strip footing supported in unreinforced, and sand reinforced with a geotextile bed subjected to static and repeated loads. Rate of settlement decreased as number of loading cycles increased and with increase in amplitude of repeated load regardless of reinforcement. Reference [5] studied foundation behavior on geosynthetic reinforced sandy soil. From the study it was concluded that with two or more reinforcement layers, settlement can be reduced to about 20%. Results also showed that sand reinforced with geogrid and geotextile composite behaved superior than sand reinforced with geogrid or geotextile alone. Reinforcement also distribute load more uniformly hence reducing stress concentration. Reference [2] studied the behavior of foundations on clayey soil reinforced with geogrid. The result showed increased bearing capacity and reduction in settlement. Settlement could be reduced by more than 50% after providing three or more reinforcing layers.

In the present study elaborated laboratory model footing tests on unreinforced and reinforced soil have been done to analysis the effect of all the parameters used in the study for the improvement of bearing capacity and settlement reduction using geogrid reinforcement. The experimental results obtained are presented and discussed in results section.

II. MATERIALS PROPERTIES

A. Clay

In the experimental work, the clay used is collected from Tekanpur town, near Gwalior, Madhya Pradesh, India. It is of intermediate plasticity having plasticity index 24. It is then cleaned to free it from debris, grass roots, other organic matter. And oven dried for 24 hours to remove moisture and passes through 75 mm IS sieve. Average unit weight of undisturbed un-situ clay was 18.7kN/m^3, having cohesion as 83kPa and coefficient of internal friction as 4°. Maximum dry density of clay was noted as 16.1kN/m^3 at an optimum moisture content of 17.2%. Properties of clay is provided in TABLE I.

TABLE I. PROPERTIES OF CLAY

Properties	Value
Specific Gravity	2.7
Percentage passing IS Sieve 75 micron (%)	83
Liquid limit (%)	43
Plastic limit (%)	19
Plasticity index (%)	24
Maximum dry Unit Weight (kN/m^3)	16.1

B. Geogrid

In the study, single type of geogrid is used. It is made from polypropylene. The properties of geogrid are provided by the manufacturer is shown in TABLE II.

TABLE II. PROPERTIES OF GEOGRID

Properties	MD	TD
Tensile Strength (kN/m)	7.8	8.2
Tensile Modulus at 0.5% strain (kN/m)	550	350
Aperture Size (mm)	30×30	
Material	Polypropylene	

III. EXPERIMENTAL SETUP

In this section, experimental procedure is described in different steps.

A. Prepration of testing bed

All the experiments were conducted in a steel tank of dimension 750 mm × 450 mm × 600 mm. Air dried clay was filled in tank in three lifts which were compacted at its maximum dry density of 16.1kN/m^3. Clay is mixed with optimum water which was determined from Standard proctor test. Soil is put in tank and compacted with help of rammer having base of 150 mm falling freely from a height between 25mm to 102mm depending upon reinforcement spacing. In the end of compaction, sprit level was used to check horizontal surface of bed.

B. Placement of Geogrid

Geogrid of required size was cut and placed in the soil at the appropriate location as calculated. After placement of geogrid, soil was again filled in the tank up to top location.

C. Equipment Setup

Test was performed in tank with one longitudinal side of tank made of acrylic sheet to get the visual observation. Square footing of dimension 75mm × 75mm was used as surface footing placed at the center of clay bed. The sides of the tank were fabricated with steel sheet to counter the bulging of the steel tank. Load was applied through Compression Testing Machine and the settlement was recorded with the help of two dial gauges attached having accuracy of 0.001 mm. Load cell of 25kN capacity was placed at center. Fig. 1 shows the schematic view of the soil reinforced with geogrid. Bearing capacity ratio (BCR) is defined as the ratio of bearing capacity of reinforced soil to that of unreinforced soil. Bearing capacity is obtained by double tangent method. Two tangents were drawn, one from the start of the curve and another from the end of the curve. The projection of interception on x-axis is the observed bearing capacity.

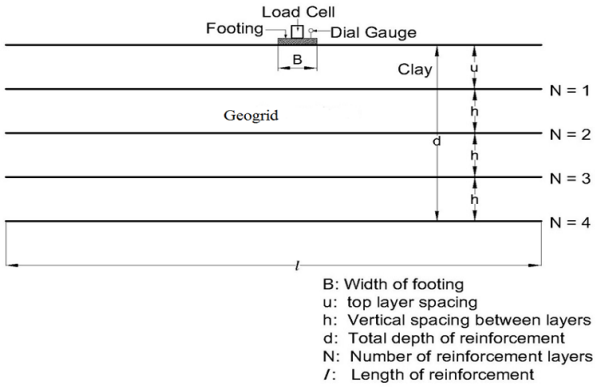


Fig. 3 Geogrid Reinforced Clay foundation

D. Testing Procedure

The objective of the study was to investigate geogrid effect on the bearing capacity of clay. The parameters considered were vertical spacing of first geogrid layer from footing base (u), effective depth of reinforcement (d), number of geogrid layers (N). At first, test on clay was performed at unreinforced condition to compare the result with reinforced condition. Load was applied at the center of the square footing with help of compression testing machine. The resultant settlement of the footing recorded with help of dial gauge. For reinforced condition, optimum vertical spacing between geogrid and footing base, 'u' was found. 'u/B' values were varied form 0.17, 0.34, 0.51, 0.68 and 0.85 respectively. Likewise, number of geogrid layers were varied from N=1 to N=5 by taking the vertical spacing between the layers of geogrid as 'u' optimum.

IV. RESULTS AND DISSCUSSION

A. Effect of first reinforcement depth (u)

The effect of first reinforcement depth i.e. u was investigated using a single layer of geogrid located below footing base. u/B is defined as the distance of the first geogrid layer from footing base (u) to the width of the footing (B). Fig. 2 shows pressure settlement curve of reinforced clay footing for single geogrid. From the graph it is observed that value of bearing pressure increases with u/B. Fig. 3 shows the relationship between bearing capacity ratio (BCR) and u/B. Fig. 3 shows variation of BCR with u/B at different settlement. From Fig. 3, is noted that BCR increases with the increase in u/B values. However, after a certain point the BCR decreases with increase in u/B. Hence, optimum value of 'u' is approximately found between 0.34B-0.51B.

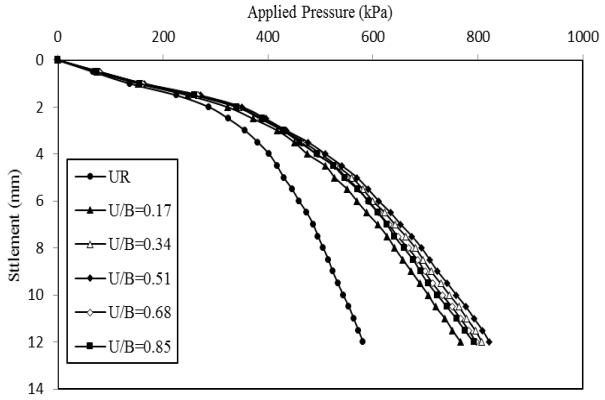


Fig.2 Pressure- settlement curves for Geogrid

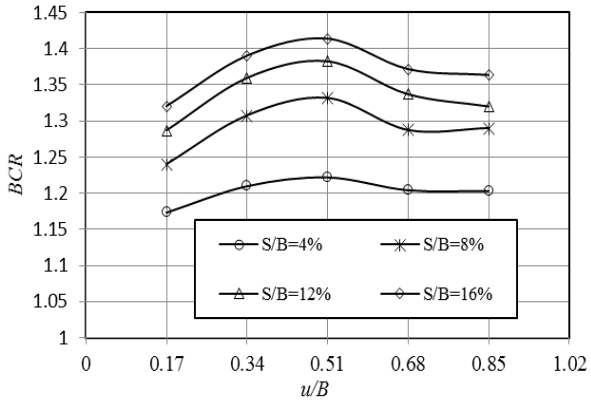


Fig. 3 Bearing capacity ratio (BCR) for geogrid for (u/B)

Reference [2] from their study found that clay when reinforced with geogrid observe maximum bearing pressure at 0.33B for single rinforcement layer. Reference [10] noted that for sand reinforced with geocell, optimum distance between footing base and first reinforcement layer falls between 0.25B-0.40B. Similar results were found by [9]. For clay reinforced with geotextile, optimum vertical spacing is between 0.35B-0.40B.

B. Optimum depth of effective reinforcement (d/B)

Effective depth of reinforcement is the depth beyond which the influence of reinforcement in the improvement of bearing capacity is almost negligible. In the study, effective reinforcement depth is denoted by'd'. From the literatures, it has been noted that'd' is equivalent to number of reinforcement layers. The relation between'd' and 'N' is shown as in (1)

$$d = u+(N-1)\times h \quad (1)$$

Where, h is the vertical spacing between two reinforcing layers, 'u' is the first reinforcement depth and N is the number of reinforcing layers. Fig. 4 shows the variation between BCR and effective reinforcement depth. From the figure it is observed that, the BCR increases with increase in the reinforcement depth is nearly equal to 1.36B. Further increase in d/B value, no significant increment in bearing capacity was observed. Therefore, optimum number of reinforcements was found equal to 4.

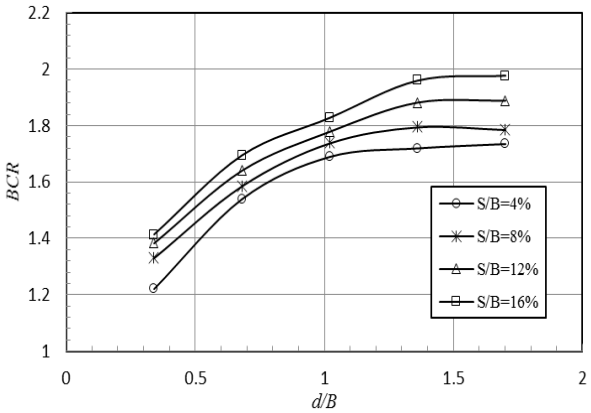


Fig. 4 Bearing capacity ratio (BCR) for geogrid for (d/B)

Reference [11] found influence depth for for strip footing over geogrid reinforced clay is 1.8B. Reference [2] for clay reinforced with geogrid found influence depth at 1.5B depth.

V. CONCLUSION

Based on tests performed, following conclusion were drawn:

- Optimum spacing between footing base and first geogrid was found in between 0.35B-0.51B for surface square footing on clay.

- Load carrying capacity of clay enhances with increase in geogrid numbers up to a certain depth. This depth is called as influence depth or effective reinforcement depth (d).
- Influence depth (d) was found nearly around 1.36B.

The study shows that geogrid can be incorporated in the clay to increase its bearing capacity hence a win-win condition for the clay having properties like that of clay used in the study.

REFERENCES

[1] J. Binquet and L. K.L, "Bearing Capacity Tests on Reinforced Earth Slabs," *Journal of Geotechnical Engineering Division, Vol. 101, No. GT12,* pp. 1257-1276, 1975.

[2] Q. Chen, M. Farsakh, R. Sharma and X. Zhang, "Laboratory Investigation of Behaviour of foundation on Geosynthetic reinforced clayey soil," *Transportation Research Record: Journal of the Transportation Research Board,* pp. 28-38, 2004.

[3] B. Das and M. Omar, "The Effects of Foundation Width on Model Tests for the Bearing Capacity of Sand with Geogrid Reinforcement," *Geotechnical and Geological Engineering, Vol. 12,* pp. 133-141, 1994.

[4] J. Mandal and H. Sah, "Bearing Cpacity Tests on Geogrid Reinforced Clay," *Geotextile and Geomembranes, Vol.11,* pp. 327-333, 1992.

[5] M. Farsakha, Q. Chen and R. Sharma, "An experimental evaluation of the behaviour of the footings on Geosynthetic reinforced sand," *The Japanese Geotechnical Society, Vol.53,* pp. 335-348, 2013.

[6] T. Ingold and K. Miller, "Analytical and Laboratory Investigation of Reinforced Clay," *Proc., 2nd International Conference on Geotextile, Vol. 3,* pp. 587-592, 1982.

[7] A. Lavasan and M. Ghazavi, "Behaviour of closely spaced square and circular footings on reinforced sand," *Soil and Foundations, Vol. 52(1),* pp. 160-167, 2012.

[8] K. Lee and V. Manjunath, "Experimental and Numerical Studies of Geosynthetics-Reinforced Sand Slopes Loaded with a Footing," *Canadian Geotechnical Journal, Vol. 37,* pp. 828-842, 2000.

[9] J. Sakti and B. Das, "Model Tests for Strip Foundation on Clay Rreinforced with Geotextile Layers," *Transportation Research Record No. 1153, National Research Council, Washington, D.C,* pp. 40-45, 1987.

[10] A. Shadmand, M. Ghazavi and N. Ganjian, "Load-Settlement characteristics of large scale square footing on sand reinforced with opening geocell reinforcement," *Geotextiles and Geomembranes, Vol.46,* pp. 319-326, 2018.

[11] E. Shin, B. Das, V. Puri, S. Yen and E. Cook, "Bearing Capacity of Strip Foundation on Geogrid-Reinforced Clay," *Geotechnical Testing Journal, Vol.16, No.4,* pp. 534-541, 1993.

[12] M. Tafreshi and A. Dawson, "Laboratory tests of footing supported on Geotextile reinforced sand under repeated loading," *The 15th World Conference on Earthquake Engineering, Vol.25,* 2012.

[13] T. Yetimoglu, T. Jonathan and A. Saglamer, "Bearing Cpacity of Rectangular Footing on Geogrid Reinforced Sand," *Journal of Geotech Engineering, ASCE, Vol.120(12),* pp. 2083-2099, 1994.

Multimedia Tools and Applications
https://doi.org/10.1007/s11042-019-07938-x

Adaptive enhancement of underwater images using multi-objective PSO

Rajni Sethi [1] · Indu Sreedevi [2]

Received: 2 May 2018 / Revised: 28 April 2019 / Accepted: 4 July 2019
Published online: 25 July 2019
© Springer Science+Business Media, LLC, part of Springer Nature 2019

Abstract

Underwater images have poor clarity and bad contrast due to low illumination in deep water. Moreover, underwater images are bluish-green in appearance due to inherent wavelength absorption property of water. Therefore, the study of underwater images is a difficult task. Being computationally simple, histogram-based enhancement techniques are obvious choice for improvement of contrast and color of underwater images. However, due to lack of any guidance mechanism, these techniques can overstretch the histogram leading to artifacts in the image. Hence, an adaptive method named 'Contrast and Information Enhancement of Underwater Images' (CIEUI) is proposed, which enhances underwater images by improving their contrast and information content using Multi-Objective Particle Swarm Optimization (MOPSO). Objective functions of MOPSO are chosen to act as guiding mechanism to ensure color & contrast correction and information enhancement respectively without introducing artifacts. Computed results not only have good contrast and color performance but also have better information content. The proposed CIEUI technique performs quantitatively and qualitatively better as compared to state-of-the-art algorithms.

Keywords Underwater Images · Image Enhancement · MOPSO · Color correction · Fuzzy logic

1 Introduction

Oceans cover more than 70% of the earth's surface and yet their environment is the least explored by humans. Capturing clear images for underwater scientific studies is a

✉ Rajni Sethi
rajni_gumber23@yahoo.com

Indu Sreedevi
sindu@dce.ac.in

[1] Department of Information Technology, Delhi Technological University, Delhi, India

[2] Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, India

very troublesome task [5]. Underwater images play a substantial part in scientific missions for various applications, e.g. assessing water quality [35] [38], observing aquatic life through various sensor data [36] [37], population tally of aquatic species and assessing geological or biological environments [33]. When the light enters water, most of it is reflected back and only a portion of it actually enters water resulting in scattering of light. Moreover, there is poor lighting during the acquisition of images; as a result, the penetrated light is further absorbed by the water. In addition, due to the inherent physical property of the water, the perceived color of the aquatic objects becomes bluish as the depth increases. The amount of color distortion is generally dependent on the depth at which image is captured, where the higher wavelengths in the visible spectrum (light in the red region) could be strongly attenuated with increasing depth, while the lower wavelengths in the visible spectrum (light in the blue region) suffer from minimal attenuation [49]. All these above factors lead to very poor contrast and color distortion in underwater images. Therefore, these images require contrast as well as color correction for underwater scientific studies. Special Optical Filters or polarizers are available for capturing underwater images [44] but such filters can lessen part of the approaching radiation and frequently change the actual shades of the scene to be captured. Artificial lighting can be used to countermeasure the bad contrast in underwater images but such a measure leads to inconsistencies in the intensity of the scene such as uneven illumination [15].

Thus, an effective solution to the problems in underwater images, should correct color distortion and contrast of the images. The image processing algorithms can either process an image using enhancement algorithm or restoration algorithm. Unlike the latter algorithms, which require parameter estimation of image formation or noise model to process the image, the former algorithms do not require any information of any parameter of the image. As image enhancement algorithms try to tweak the histograms of the image to improve its color and contrast, it may lead to overstretching of the histogram since there is no guiding mechanism. Since the underwater medium can also vary, i.e., salinity, turbidity, organic matter of the medium can also affect the image quality [39], there is need of a method which can handle both color and contrast problem in different underwater media. In literature, both image enhancement and restoration based techniques have been employed to improve the quality of underwater images, but owing to simplicity, enhancement based algorithms is preferred in the proposed work.

In this paper, a method called ‘Contrast and Information Enhancement for Underwater Images’ (CIEUI) has been formulated for this purpose. CIEUI employs Multi-objective Particle Swarm Optimization (MOPSO) to perform contrast correction and information enhancement simultaneously. The parameters of CIEUI decide the degree of color correction for different media which are tuned by MOPSO. Objective functions of MOPSO are the performance measures which act as a guiding mechanism so as to control the overstretching of histogram. CIEUI delivers results with enhanced information and contrast as compared to other algorithms for underwater images.

The remaining part of this paper is organized in the following order. A brief literature review for the underwater image enhancement is presented in section 2. CIEUI is explained in section 3, in which the fuzzy gray world technique from [45], performance measures, MOPSO and its parameters are elaborated on. A performance comparison of CIEUI with the other related algorithms is shown in section 4. Finally, the presented technique is concluded in section 5.

2 Background and motivation

2.1 Overview

Underwater images are of poor quality with bluish-green (color cast) and washed out (haze) appearance with poor contrast. In literature, the research work is oriented either towards removing color cast or removing haze from the images. A lot of work has been done on underwater images by focusing on their histogram as it does not require the knowledge of image formation and concentrates only on improving the visual quality of the image by applying operations in the spatial domain. Generally, in underwater images, the histogram of the red channel is skewed towards lower intensities, the histogram of green channel usually covers the whole dynamic range and the histogram of blue channel is skewed towards higher intensities. Histogram-based methods use qualitative measures to deliver an image which is visually appealing and improve the contrast of underwater images by stretching the histograms of the red and blue channel to cover all intensities.

2.2 State-of-the-art methods

Automatic Color Equalization (ACE) is an unsupervised color equalization algorithm proposed in [43], based on adaptation mechanisms (e.g. color constancy, lightness constancy) of human vision system. ACE was applied on underwater videos by a color correction method [8]. But tuning of inner parameters of ACE was required to get the histogram shape closer to naturalness for enhancing underwater images. Unsupervised Color Correction Method (UCM) [25] and integrated color model (ICM) [26] are the techniques for underwater images and are based on gray world assumption [6] and white patch. Although, both the techniques increase the overall contrast of the image, but results appear almost same. The output images are under-enhanced in some areas making these areas too dark. Moreover, these techniques tend to produce images with high-noise. Ghani *et al.* [17] modified and extended UCM and ICM to overcome the shortcomings of these methods. It modifies the stretching process of histogram in such a way that a bell-shaped distribution of the image is obtained. The histogram of the output image thereby follows the Rayleigh distribution.

All the above methods follow gray world assumption and give satisfactory results. On the other hand, Henke *et al.* [20] formulated a new hypothesis in the context of underwater images, stating that the general color constancy hypothesis of achromatic average is not applicable for removal of color cast in underwater images but the results are not satisfactory. This makes it clear that the gray world assumption can be applied for color correction of underwater images.

Another method called mixture contrast limited adaptive histogram equalization (CLAHE-Mix) is proposed by Hitam *et al.* [21] for underwater images. This method extends CLAHE [55] for underwater images to reduce the noise, but in some cases, output images have higher noise as compared the conventional CLAHE. Another histogram based method proposed by Singh *et al.* [47] applied recursive histogram equalization dependent on exposure values for image enhancement which gives significant results for low-light images. [1] proposed RAHIM (Recursive Adaptive Histogram Modification) which processed histograms column-wise to enhance underwater images. This method produced enhanced images with good information content but contrast is poor. [16] performed underwater image enhancement by blending CLAHE and percentile methodologies but the color quality of the images is poor. Ancuti

et al. [4] proposed an image enhancement algorithm which fuses the two versions of a single image to address color and contrast of the underwater images but it does not work for every type of underwater image captured in different media (different turbidity, salinity).

Fuzzy logic [54] has been employed for the color correction of images as well e.g. A cost-effective fuzzy inference model based on tree structured fuzzy rules has been proposed for color correction and image enhancement. Another fuzzy inference system proposed by Hung *et al.* [23] forecasts the color cast based on which image is adjusted from the unknown illuminant to the image in standard illuminant.

A combination of evolutionary techniques with fuzzy technique gives better image enhancement, as claimed by a few researchers. Research has been done for non-aquatic images using fuzzy and evolutionary techniques, e.g. a fuzzy system for image enhancement proposed by Handmandlu *et al.* [18] used Bacterial Foraging Optimization (BFO) to determine the parameter of membership function for optimizing visual factors and entropy. Another image quality enhancement technique proposed by Verma *et al.* [52] improves the information factor by employing Artificial Ant Colony System (ACS). Kwok *et al.* technique [31] employed MOPSO for color correction and enhancement. Another technique based on MOPSO is proposed by Kanmani *et al.* [29] which works for contrast enhancement of natural images. However, all the above mentioned techniques for image enhancement do not give satisfactory results, when applied on underwater images. A PSO based technique is proposed for underwater images by Abunaser *et al.* [2] but it requires repeated execution of PSO to tune the parameters for image enhancement.

In recent years, the focus of research has shifted towards haze removal in underwater images using dark channel prior [19]. Chao *et al.* [9] and Carlevaris-Bianco *et al.* [7] used dark channel prior to clarify the blurred underwater images but failed to color correct the images. [53] also employed dark channel prior for image deblurring followed by a color correction method. Chaing *et al.* [10] proposed a dark channel prior based algorithm which further removes artifacts caused by artificial light. [28] developed a dark channel prior based algorithm in which the estimation of dark channel is done using image blurriness model. [13] formulated the problem around red channel restoration as red color attenuates the most in underwater environment. [14] modified the dark channel prior to bright channel prior to enhance the underwater images. [34] proposed an algorithm which used minimum information loss principle for image dehazing followed by a contrast enhancement algorithm for further enhancement of the image. Peng *et al.* [41] proposed a method for estimating the depth of the image formation model based on image blurriness and light absorption pattern. These dark channel prior based methods give good results which are highly dependent on correct estimation of transmission map. Moreover, there are computationally expensive. So, these methods work only for a few underwater images.

2.3 Motivation

As discussed earlier, dehazing based methods are computation intensive and depend highly on the correct estimation of depth map. Thus, the proposed work focuses on image enhancement method based on histogram modification. The histogram-based methods work well for underwater images but overstretch the histogram thereby introducing artifacts (e.g. false colors) in the image since there is no guiding mechanism to control the extent of histogram stretching. These artifacts degrade the information content of the image. However, this shortcoming can be avoided by optimizing the

information content of the image while stretching the histogram. Sethi *et al.* [45] proposed adaptive color correction for enhancing underwater images using bacterial foraging optimization for optimal color and contrast. However, output images have false colors due to overstretching of the histogram as the performance measures concentrate only on the contrast of the image. Thus, it lacks the proper guiding mechanism.

Thus, in this paper, a multi-objective optimization problem named 'Contrast and Information Enhancement for Underwater Images' (CIEUI) is formulated for adaptive color correction and information enhancement to deal with artifacts caused by overstretching of the histogram. CIEUI employs MOPSO to perform contrast correction and information enhancement simultaneously by optimizing performance measures namely 'Histogram Spread' and 'Entropy' respectively. Histogram spread ensures contrast and color performance of the image while entropy ensures information enhancement in the image without introducing artifacts. CIEUI produces better results as compared to other algorithms for underwater images. Working of CIEUI has been explained in the next section.

3 Contrast and information enhancement of underwater images (CIEUI)

In this paper, a multi- objective adaptive color correction and information enhancement technique for underwater images has been framed using the fuzzy gray world technique. CIEUI employs MOPSO to ensure that desired results are produced without artifacts like false colors, light image, halos etc. Performance measures namely 'Histogram Spread (HS)' and 'Entropy' are chosen as objective functions of MOPSO to produce an output image with enhanced contrast and information content. The fuzzy gray world method is explained below:

3.1 Fuzzy gray world color correction method

Most of the research in the field of underwater image enhancement assumes that the histogram of the red channel in an underwater image is skewed towards lower intensities and the histogram of the blue channel is skewed towards higher intensities. However, use of artificial light while capturing photographs, can lead to a change in this pattern as artificial light compensates for the loss of energy of red wavelength of light on penetrating water. Figure 1 shows the image of a scuba diver (without artificial light) and a fish (captured using artificial light) with their histograms. The fish image fails the above assumption as the blue color histogram is not skewed towards higher intensities.

The fuzzy gray world algorithm does not consider the assumption that blue color is always concentrated towards higher intensities as the underwater images can be captured in different types of media having varying salinity, turbidity, organic concentration etc. which can affect the color cast pattern. So, instead of assuming color cast as blue like other conventional histogram based techniques, it finds the actual color cast and performs color correction. Pseudo-code of fuzzy gray world based color correction is shown in Algorithm 1 and is explained below.

Consider the RGB image I_{ij} where i and j represent the pixel coordinate $I_{ij} = (R_{ij}, G_{ij}, B_{ij})$ with $R_{ij}, G_{ij}, B_{ij} \in [0,255]$. The first step is to calculate of the mean intensities of the red, green and blue channels of the given Image I. Means are denoted by $\overline{R}, \overline{G}$ and B respectively and are calculated by using Eq. 1.

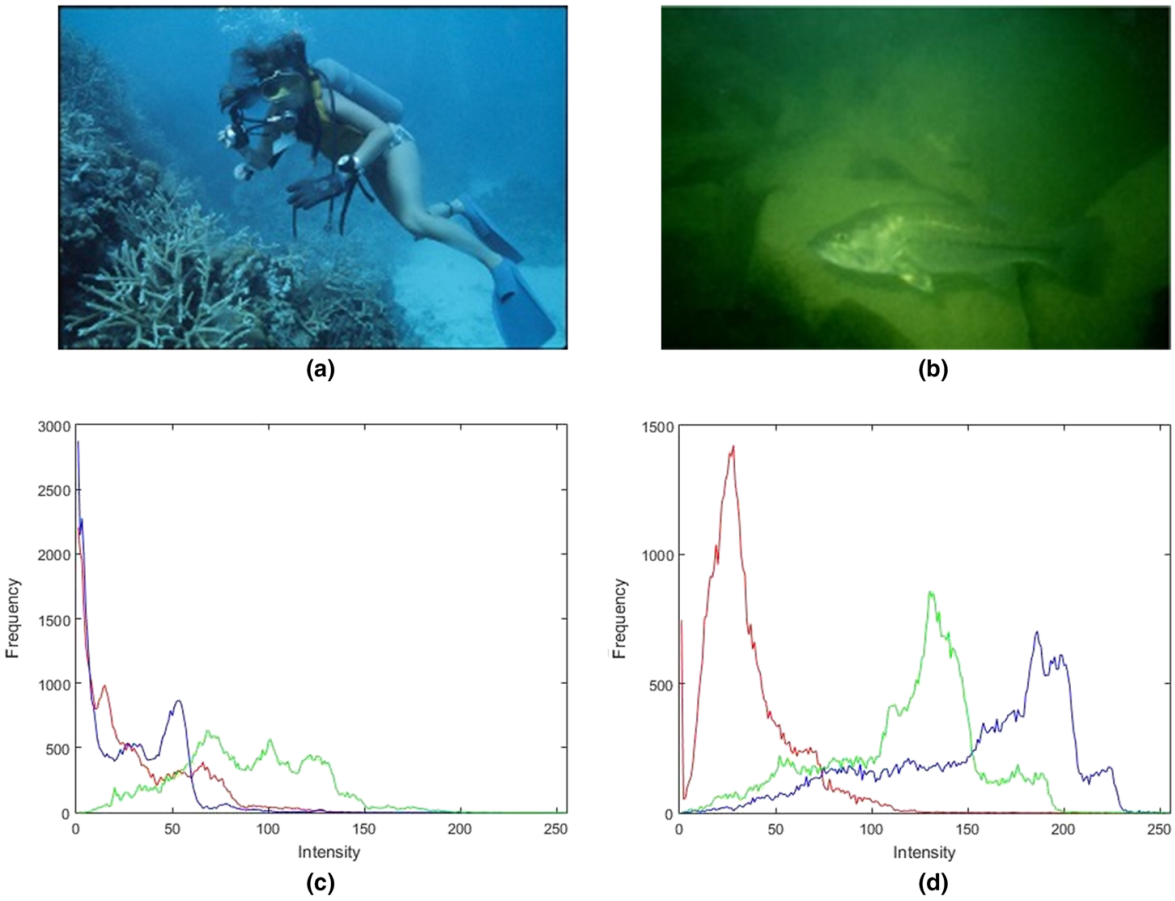


Fig. 1 **a** Scuba Diver **b** Fish **c** Histogram of the Scuba Diver image **d** Histogram of the Fish image

$$\overline{R} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}R_{ij} \quad \overline{G} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}G_{ij} \quad \overline{B} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}B_{ij} \tag{1}$$

The second step is to fuzzify the individual color channels. There are two fuzzy sets named 'LOW' and 'HIGH' which indicate the extent to which an intensity value is low and high respectively. These fuzzy sets use the gaussian membership function given by Eq. 2 and curves for the fuzzy sets are shown in Fig. 2.

$$\mu(x) = \exp\left(\frac{-(x-c)^2}{2(\sigma^2)}\right) \tag{2}$$

The parameter 'c' is the center of the gaussian curve at which membership function will have value '1'. The parameter σ is the standard deviation of the gaussian curve and defines the width of the curve.

The membership functions for the two fuzzy sets, namely LOW and HIGH are stated in Eq. 3.

$$\mu_{LOW}(x) = \exp\left(\frac{-x^2}{2(35^2)}\right) \quad \mu_{HIGH}(x) = \exp\left(\frac{-(x-255)^2}{2(35^2)}\right) \tag{3}$$

$\mu_{LOW}(x)$ and $\mu_{HIGH}(x)$ are membership value for intensity value x in the fuzzy sets LOW and HIGH respectively. Thus, the center of the curve is chosen as '0' for 'LOW' fuzzy set because

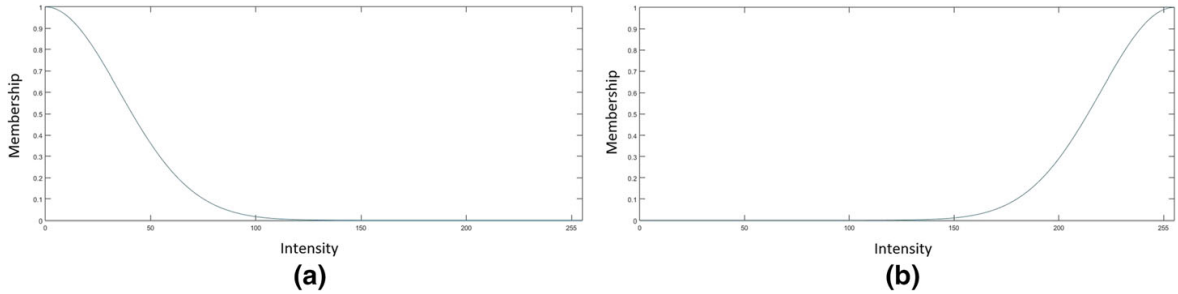


Fig. 2 Gaussian Membership for fuzzy sets (a) 'LOW' and (b) 'HIGH'

intensity 0 is the lowest possible intensity value and should have membership value as 1 for LOW. Similarly, the center of the curve for HIGH fuzzy set is chosen as 255 which is the highest possible intensity value. Value of σ for the fuzzy sets should be such that the gaussian curve follows the following conditions:-

i) At 0, Membership value in LOW fuzzy set should be 1 and start decreasing gradually with increasing intensity values. For intensities above 128, membership value should become '0' as these intensities are no longer low.
ii) At 255, Membership value in HIGH fuzzy set should be 1 and start decreasing gradually with decreasing intensity values. For intensities below 128, membership value should become '0' as these intensities are no longer high.

These conditions should be followed by every image. Value of σ is found experimentally as '35' as it is fulfilling the above-mentioned criteria. Membership of each pixel in the individual channel is found by using Eq. 3 and the pixel will belong to the fuzzy set with high membership value.

Table 1 states six fuzzy rules and base for color correction. These rules are applied to every pixel in the three channels and the rule which is followed by the maximum number of pixels decides bthe color cast of the image. Color cast gives the base for subsequent color correction. Base is a matrix of size same as that of the image. It helps in computing the scale factors for each pixel in the three color channels.

The next step is to find scale factors by using base decided from Table 1 and means calculated from Eq. 1. Scale factors α_r, α_g*and* α_b are calculated using Eq. 4.

Table 1 Fuzzy rules and base channel

Rule Number	R(i,j)	G(i,j)	B(i,j)	Cast	Base
1	High	Low	Low	Red	$\overline{R}$
2	Low	High	Low	Green	$\overline{G}$
3	Low	Low	High	Blue	$\overline{B}$
4	High	High	Low	Red and Green	$\frac{(R+G)}{2}$
5	Low	High	High	Green and Blue	$\frac{(G+B)}{2}$
6	High	Low	High	Red and Blue	$\frac{(\overline{R}+G)}{2}$

$$\alpha_r = \frac{Base}{R} \alpha_g = \frac{Base}{G} \alpha_b = \frac{Base}{B} \tag{4}$$

The final step is to perform color correction by using Eq. 5. Color correction gives corrected values for red, green and blue color channels denoted by $\hat{R}$, $\hat{G}$ and $\hat{B}$ respectively. λ_r, λ_g *and* λ_b are non-linearity factors for red, green and blue color channels respectively. These factors control the degree of color correction of each color channel. Since underwater images can be captured at different depth, in different salinity and turbidity levels of water etc., the degree of color correction can't be same for every image. This degree is tuned using Multi-objective Particle Swarm Optimization (MOPSO).

$$\hat{R} = {\alpha_r}^{\lambda_r} R \;\; \hat{G} = {\alpha_g}^{\lambda_g} G \;\; \hat{B} = {\alpha_b}^{\lambda_b} B \tag{5}$$

Algorithm 1: Fuzzy Gray World Based Color Correction

Input: RGB Image I,λ_r, λ_g *and* λ_b
Output: HS, Entropy
Result: ColorCorrectedImage

1 Separate R,G,B channels of Image;
2 [m; n] is size of each channel;
3 Compute Channel means using Eq. 1;
4 Initialize Rulecount[6] to zero for i = 1 to m do
5 for j = 1 to n do
6 μ_{LOW}(R(i; j)) and μ_{HIGH}(R(i; j)) using Eq. 3;
7 μ_{LOW} (G(i; j)) and μ_{HIGH}(G(i; j)) using Eq. 3;
8 μ_{LOW} (B(i; j)) and μ_{HIGH}(B(i; j)) using Eq. 3;
9 Find the Fuzzy Membership of pixel in each channel using Mamdani Operator;
// Mamdani Operator assign the pixel to the fuzzy set which has higher membership value
10 k = fuzzy rule followed by the pixel in Table 1 ;
// value of k ϵ [1-6]
12 Rulecount[k] = Rulecount[k] + 1;
13 end
14 end
15 [max, DominatingRuleNo] = Maximum(RuleCount);
16 Find the base from Table 1 corresponding to DominatingRuleNo ;
17 Compute $\hat{R}$, $\hat{G}$, $\hat{B}$ using Eq. 5
18. ColorCorrectedImage = [$\hat{R}$, $\hat{G}$, $\hat{B}$];
19. Compute Histogram Spread and Entropy of the ColorCorrectedImage.

Values of these non-linearity factors λ_r, λ_g *and* λ_b are computed using MOPSO algorithm which optimizes performance measures namely histogram spread [51] and entropy explained in the next section.

3.2 Performance measures

Since underwater images have diminished color, contrast and bad illumination, performance measures chosen should ensure that color, contrast and information should improve. Two metrics chosen as performance measures are namely Histogram Spread (HS) [51] and Entropy [31]. HS [51] is a quantitative measure for assessing the overall contrast of an image. It quantifies the uniformity of histogram over the entire intensity range and should be close to 0.5 if the histogram is uniformly spread. Since for a good contrast image, the histogram should be uniformly spread across the whole intensity range, HS

gives an index for measuring the uniformity of the histogram. Entropy ensures that the image has good information content and it should not decrease after enhancement.

3.3 Multi-objective particle swarm optimization for contrast and information enhancement

There are a lot of meta-heuristic techniques proposed in the literature for multi-objective problems, e.g., non-dominated sorting genetic algorithm [48], pareto-archived evolution strategy [30] etc. These techniques are nature-inspired and evolutionary in nature. Another technique which is an advancement over these extensively used techniques is an optimization algorithm called Particle Swarm Optimization (PSO) [27]. PSO is gaining popularity due to its simplicity. The concept of population and a measure of performance used in PSO is analogous to population and the fitness value used in evolutionary algorithms. All the calculations performed in PSO are similar to crossover operator. However, there is an added advantage of solutions being scattered over a hyperspace which is not available in traditional evolutionary algorithms. Another advantage of PSO over evolutionary algorithm is that it learns from the past experience. Thus, PSO has been choice for multi-objective problem which provides better convergence rate as well.

MOPSO [11] is very similar to PSO as information about global and personal best locations is shared within flock in both the algorithms. However, PSO is slightly modified to handle multiple objectives in such a way that it discovers a set of solutions called 'Pareto Front' rather than a single global best solution. A pareto front (repository) is a set of all the non-dominated solutions found after each iteration. The detailed steps of MOPSO are shown in Algorithm 2.

MOPSO algorithm starts with the initialization of particles in the swarm with random values and various other parameters whose values are stated in Table 2. To achieve accelerated convergence and to reduce computational complexity, the number of iterations and swarm size are kept in a moderate range. Inertia weight is the control parameter for swarm velocity [50] and is given moderate value so that convergence does not become too fast without the exploration of a good range of positions. Personal learning coefficient is given lesser weight as compared to global learning coefficient so that position should be dependent more on global best positions than local best positions. Initial values of parameters have been taken by referring to state-of-the-art techniques [3] [42] and then updated after repeated executions for faster convergence and reduced time complexity.

After initialization, the first step is to find the value of performance measures for the initial population in order to find the global and personal best solutions. Thus, HS and entropy is

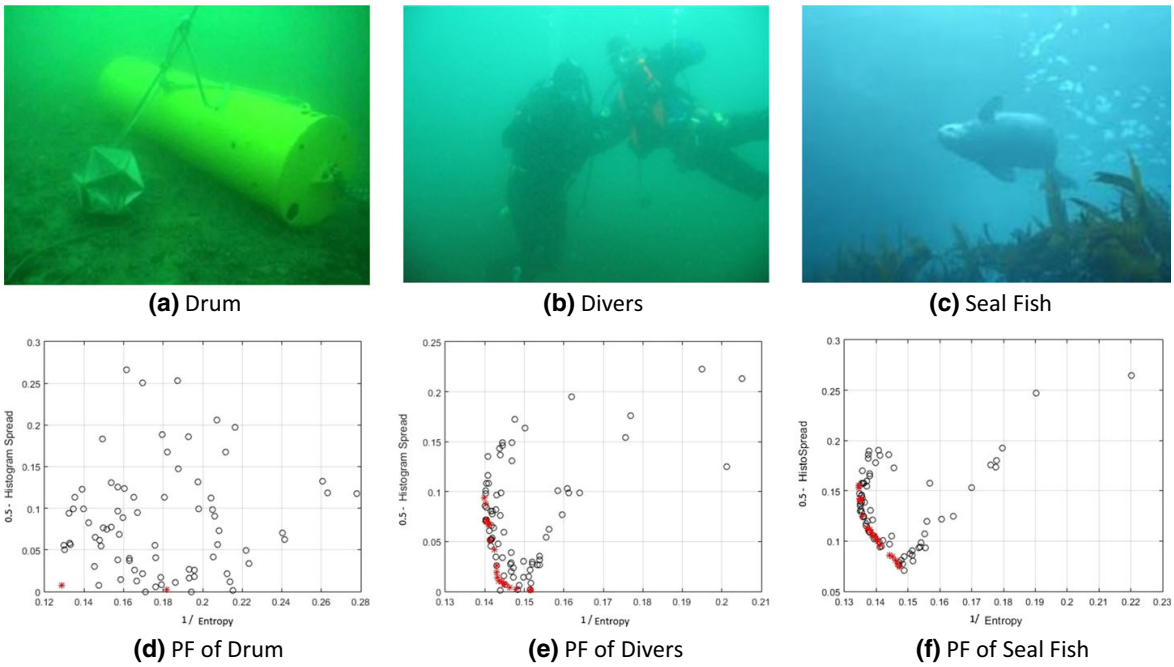
Table 2 MOPSO parameters

Parameter	Value
Number of Iterations	30
Size of Swarm	75
Size of Repository	20
Interia Weight	0.5
Personal Learning Coefficient	1.5
Global Learning Coefficient	2.5

calculated for each particle of the swarm. Then, the non-dominated solutions from the swarm are found using the pareto-optimality definitions [42] and stored in a repository.

Input: RGB Image
Output: Repository of Non-Dominated values of
Result: Enhanced Image
1. Initialize the MOPSO parameters
2. EVALUATE the value of HS and Entropy of each particle of the swarm using Algorithm 1.
3. Repository = SELECT the non-dominated solutions from the Swarm.
4. for t = 1 to MaxIterations do
5. for i = 1 to size of swarm do
6. SELECT the gbest ;
7. UPDATE the velocity ;
8. UPDATE the Position ;
9. EVALUATE the Particle again by computing HS and Entropy using Algorithm 1.
10. UPDATE the pbest;
11. End
12. UPDATE the repository with gbests;
13. Report Results in the repository;
14. end
15. Choose the particle from repository with best UICM Value
16. Output the Enhanced Image using Algorithm 1 for the chosen particle.

Each iteration finds the global best solution; updates velocities and position of particles for the swarm, just like PSO. Particles are then evaluated based on the performance measures and the personal best is found out. The repository is updated with global bests. After the end of iterations, the repository or pareto-front is generated with the global best solutions from which one solution is chosen and an enhanced image is produced using the Algorithm 1. Thus, CIEUI aims to find the optimal value of the non-linearity factors λ_r, λ_g *and* λ_b while optimizing the values of HS and entropy such that the HS value is close to 0.5 and the entropy value of the enhanced image is greater than or equal to the original image. MOPSO will provide a pareto front which is a set of non-dominated solutions which are chosen as optimal if entropy can be improved without sacrificing histogram spread and vice-versa. Pareto fronts for a few images are shown in Fig. 3. Red dots in the plots are members of the pareto front. Since there are a



(a) Drum **(b)** Divers **(c)** Seal Fish

(d) PF of Drum **(e)** PF of Divers **(f)** PF of Seal Fish

Fig. 3 Images and their Pareto-Fronts (PFs)

number of non-dominated solutions, there should be a method to choose one solution out of them. For this purpose, another performance metric named 'UICM' (Underwater Image Colorfulness measure) [40] has been considered to choose from the set of candidate solutions. The solution with the best UICM value will be chosen as the optimal values of non-linearity factors required for computing the enhanced image.

4 Results and discussion

CIEUI has been implemented using MATLAB version 2015a on Intel Core i5 at 2.60 GHz. The following methods have been used in our experiments: gray world technique [32], UCM

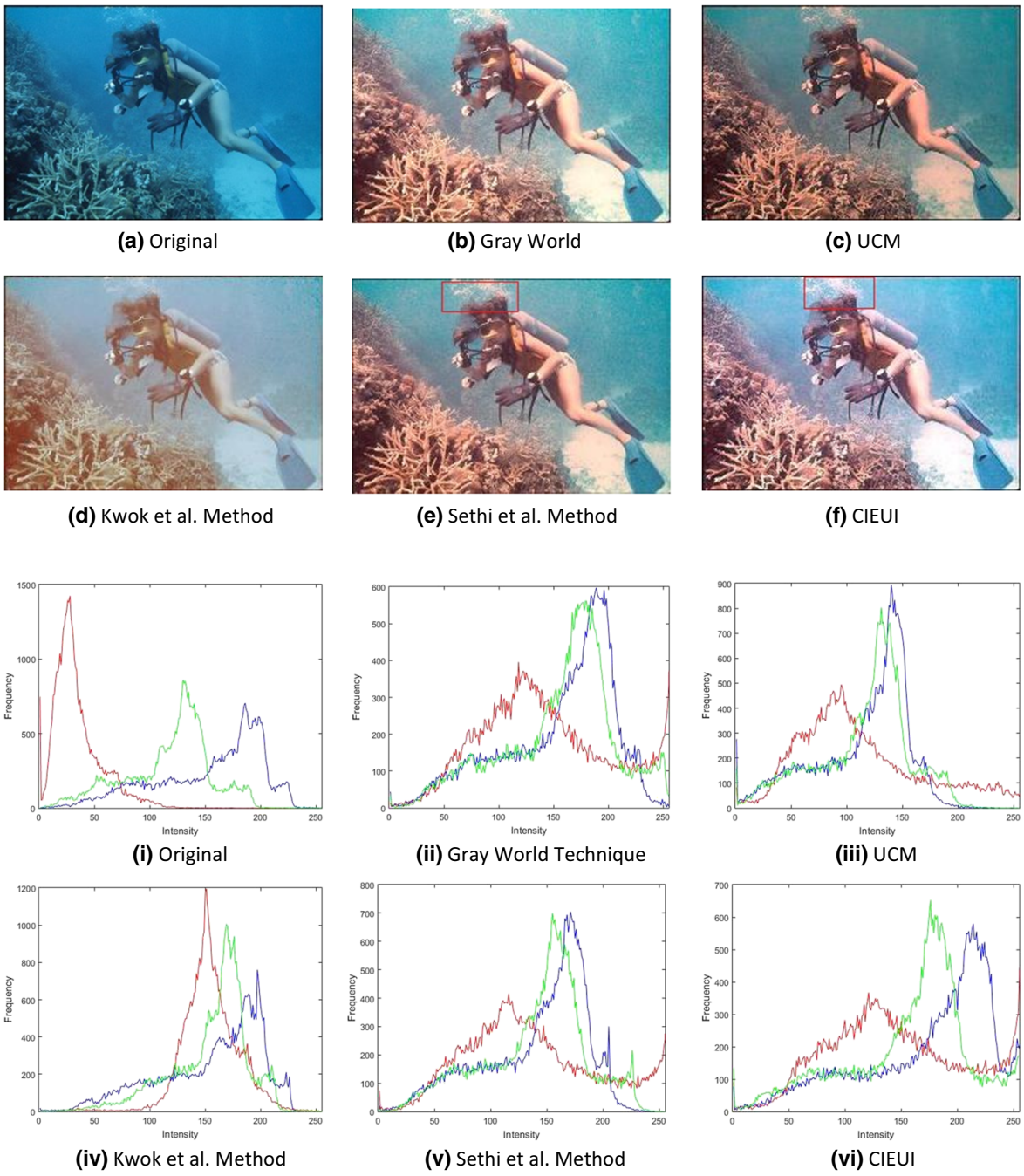


Fig. 4 (a)-(f) Results of Scuba Diver Image (i)-(vi) Histograms of the Scuba Diver Image

[24], Kwok *et al.* technique [31] and Sethi *et al.* technique [45]. To demonstrate the efficacy of CIEUI, output and histograms of a few underwater images using the above-mentioned techniques are shown in Figs. 4, 5, 6, 7 and 8. These underwater images are chosen as they are captured at different depths and in different water bodies.

Recently, a few underwater image databases have been developed namely TURBID dataset [12] and WHOI color correction dataset (Singh, [46]). The pictures of TURBID dataset are generated by experimentation for simulation of images in the turbid water body and WHOI (Woods Hole Oceanographic Institution) color correction dataset has original images along with reference images generated by the color correction method developed by scientists of WHOI. Thus, a comparison can be done with the results given on the WHOI website (H. Singh) using reference based performance measures like PSNR (Peak

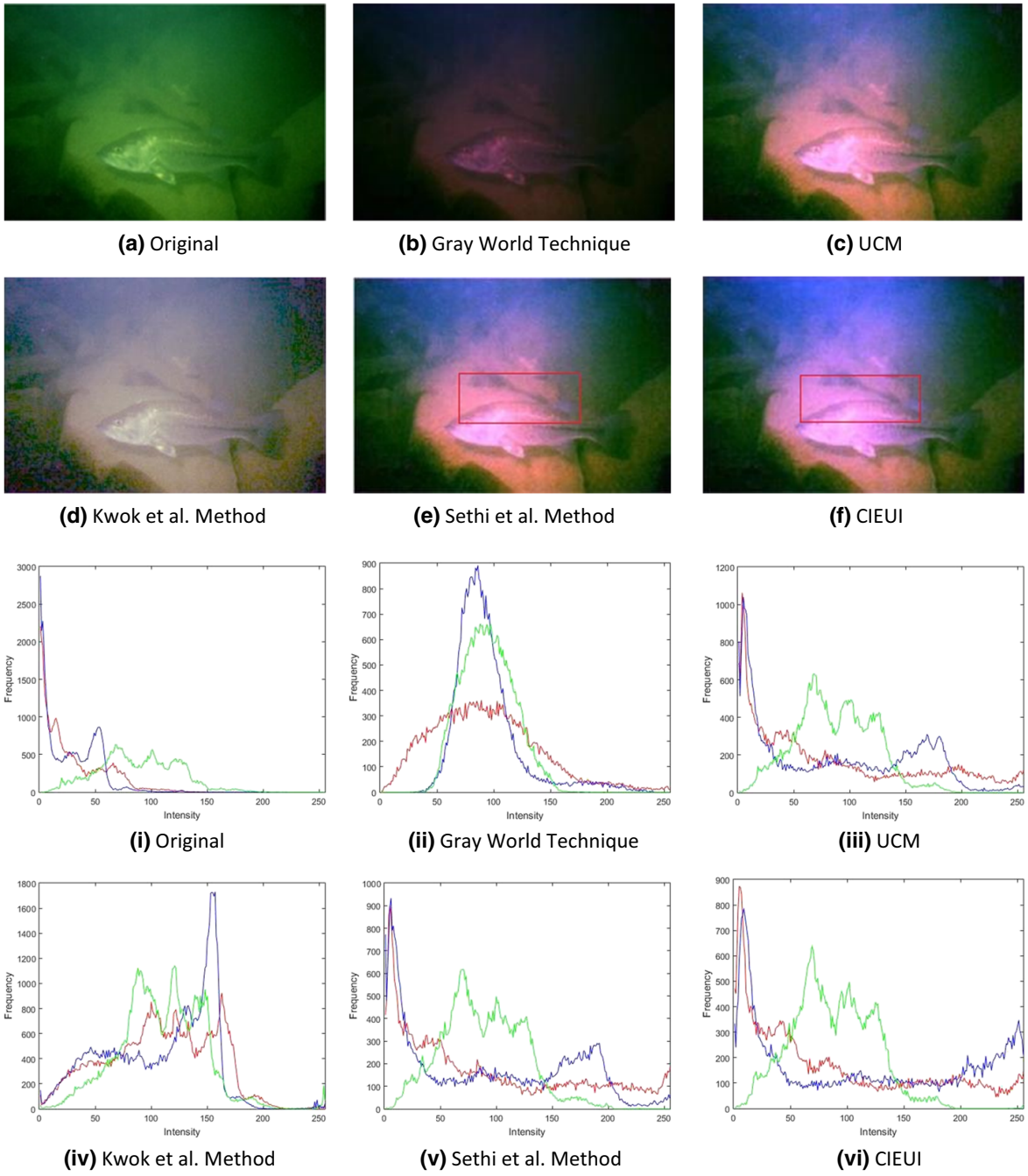


Fig. 5 (a) –(f) Results of Fish Image (i)-(vi) Histograms of the Fish Image

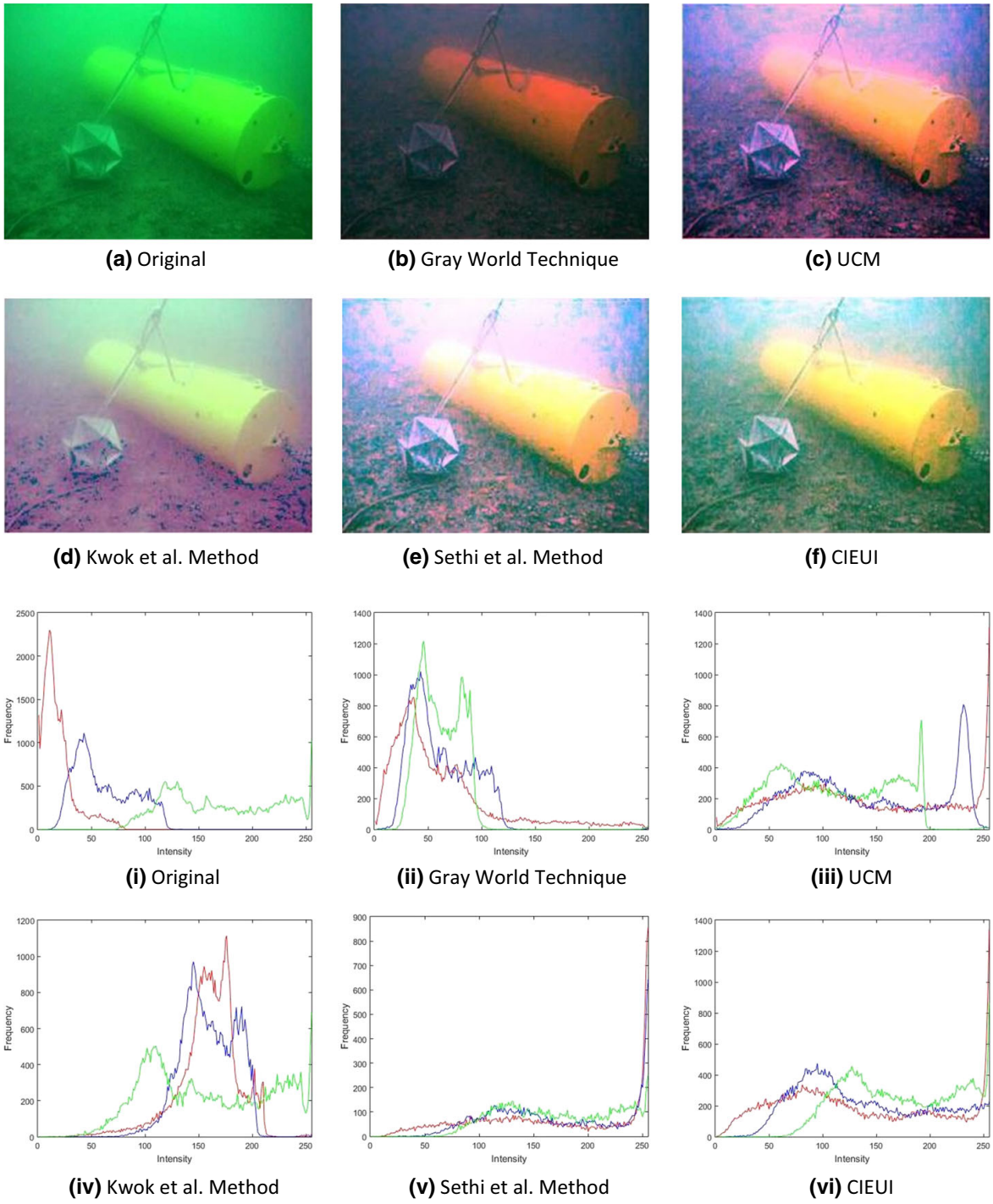


Fig. 6 (a)-(f) Results of Drum Image (i)-(vi) Histograms of Drum Image

Signal to Noise Ratio) and MSE (Mean Squared Error) but these reference based performance measures cannot be used for quantitative evaluation of other underwater images. Other than these datasets, underwater images neither have any database with ground truth values nor any universally accepted quantitative measure for evaluating the extent of image enhancement. Thus, assessing the performance of underwater image enhancement methods is very challenging. Thus, CIEUI has been tested for 200 real underwater images taken from different underwater collections (Hope [22]) and articles related to underwater image processing available on the internet for exhaustive comparison.

For quantitative analysis of the above-mentioned techniques, the values of performance measures namely 'Entropy', 'Histogram Spread' and 'UICM' have been measured.

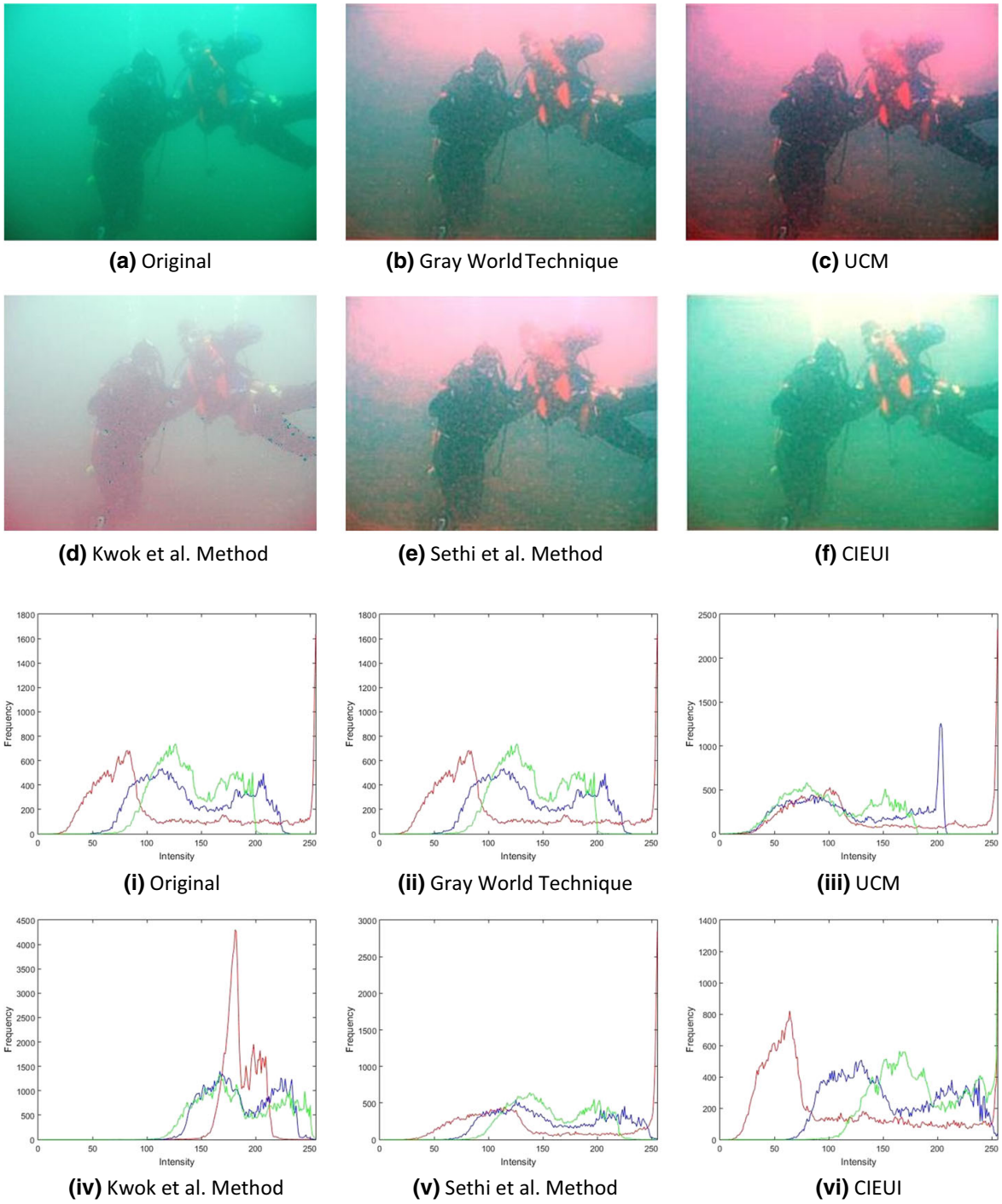


Fig. 7 (a)-(f) Results of the Divers Image (i)- (vi) Histograms of the Divers Image.

Entropy measures the information content of the image. Histogram Spread quantifies the contrast of the image. UICM measures the colorfulness of the image. The averages of the above-mentioned performance metrics for 200 underwater images for the compared techniques are shown in Table 3. The values of these performance measures for different images are listed in Table 4 to compare CIEUI with the above-mentioned techniques. The bold values in each column of Table 4 represent the best value of the respective performance measure for particular image.

From the qualitative and quantitative analysis of results of the techniques to be compared, the following points can be inferred. Original underwater images have a

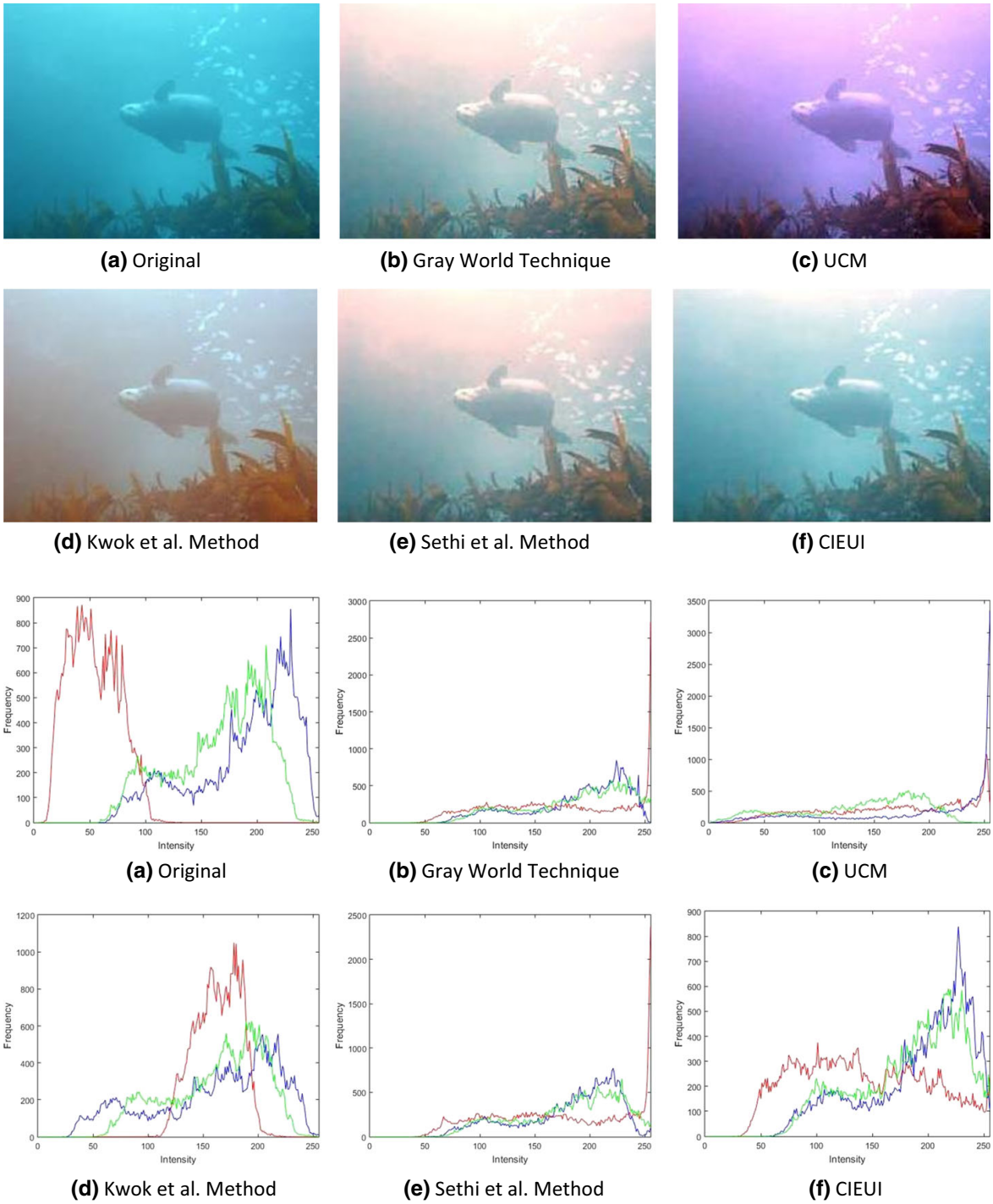


Fig. 8 (a)-(f) Results of the Seal Fish Image (i)-(vi) Histograms of the Seal Fish Image.

Table 3 Average values of Entropy, HS and UICM for 200 Underwater Images

Technique	Entropy	Histogram Spread	UICM
Original	6.05	0.2531	3.0431
Gray World	6.89	0.3238	5.8244
UCM	7.18	**0.3771**	6.9969
Kwok *et al.* Method	6.06	0.2327	3.8461
Sethi *et al.* Method	6.89	0.3615	6.2657
CIEUI	**7.52**	0.3589	**7.1539**

washed out appearance with very less information content (low entropy values). Gray world technique [32] and UCM [25] try to bring color information in the image by manipulating the histograms of individual RGB channel but these techniques sometimes overstretch the histogram leading to false colors in image. From the Table 4, it can be concluded that the gray world technique [32] and UCM [25] produce images with improved entropy (information), histogram spread (contrast) and UICM (color). But, it can be seen in the visual results of gray world and UCM techniques that red color dominates the image.

Qualitative and quantitative analysis of Kwok *et al.* [31] method proves that it improves entropy (information content) to some extent but the contrast of the images is very poor. Histogram spread values of some of its output images are sometimes lower than the original images. Thus, It does not work well for underwater images but it claims to work for poor contrast images.

The Sethi *et al.* [45] technique improves entropy (information content), histogram spread (contrast) and UICM (color), but introduces false colors in the image just like UCM and gray world technique. Entropy of its output images is not increased as much as

Table 4 Comparison of various techniques

Image	Technique	Performance Measures		
		Entropy	Histogram Spread	UICM
Drum	Original	5.65	0.4005	2.9495
	Gray World	7.10	0.2822	4.9339
	UCM	7.45	0.5577	9.5244
	Kwok *et al.* Method	6.34	0.3163	3.0975
	Sethi *et al.* Method	6.85	0.5488	5.1326
	CIEUI	**7.78**	**0.4586**	**9.7180**
Divers	Original	5.44	0.2817	4.7572
	Gray World	6.97	0.4241	4.2785
	UCM	6.59	0.5563	4.7077
	Kwok *et al.* Method	5.74	0.2417	4.9282
	Sethi *et al.* Method	6.77	**0.5097**	6.2125
	CIEUI	**7.37**	0.4078	**11.3503**
Fish	Original	6.01	0.2553	1.3890
	Gray World	5.91	0.2639	2.0289
	UCM	**7.41**	0.4333	5.8051
	Kwok *et al.* Method	7.02	0.2939	1.4890
	Sethi *et al.* Method	**7.41**	0.4446	7.0389
	CIEUI	7.32	**0.4706**	**11.4956**
Seal Fish	Original	6.46	0.3081	1.2937
	Gray World	6.86	**0.4148**	**5.1796**
	UCM	7.19	0.4003	4.6775
	Kwok *et al.* Method	6.47	0.3120	1.7709
	Sethi *et al.* Method	6.95	0.3962	4.8969
	CIEUI	**7.63**	0.3506	2.7785
Scuba Diver	Original	6.25	0.1992	1.3890
	Gray World	7.58	0.3154	2.0289
	UCM	7.53	0.2550	5.8051
	Kwok *et al.* Method	6.36	0.1804	1.4891
	Sethi *et al.* Method	7.52	0.2901	7.0389
	CIEUI	**7.59**	**0.3351**	**11.4856**

its histogram spread and UICM values. Reason for such results is that it concentrates only on contrast and color improvement which leads to introduction of false colors in the image.

Results of CIEUI technique have improved contrast and information content as compared to other techniques. In the fish image, it is introducing false colors which are due to presence of artificial light. From Table 3, CIEUI shows better overall performance than others in terms of improving entropy and UICM values. However, it lags behind UCM and Sethi *et al.* [45] technique in terms of improving histogram spread as it computes an optimal value of histogram spread and entropy for producing an image with good contrast and information content. It delivers good results for every type of underwater image. For example, drum image is captured on the seabed which is very deep, the Divers image is at a moderate depth and moreover images are taken in different water bodies having different salinity and turbidity.

The following points can be inferred from the histograms of the underwater images produced by the above-mentioned techniques and CIEUI:

Histograms of images produced by gray world based technique [32] are either shifted towards lower intensities (low contrast) or towards higher intensities (high contrast). For the scuba diver image, histograms of three color channels span the whole dynamic range but the distribution of pixels is non-uniform.

Histograms of UCM [25] images span entire dynamic range but do not have uniform distribution and the histogram of blue channel usually shows a dip for higher intensity

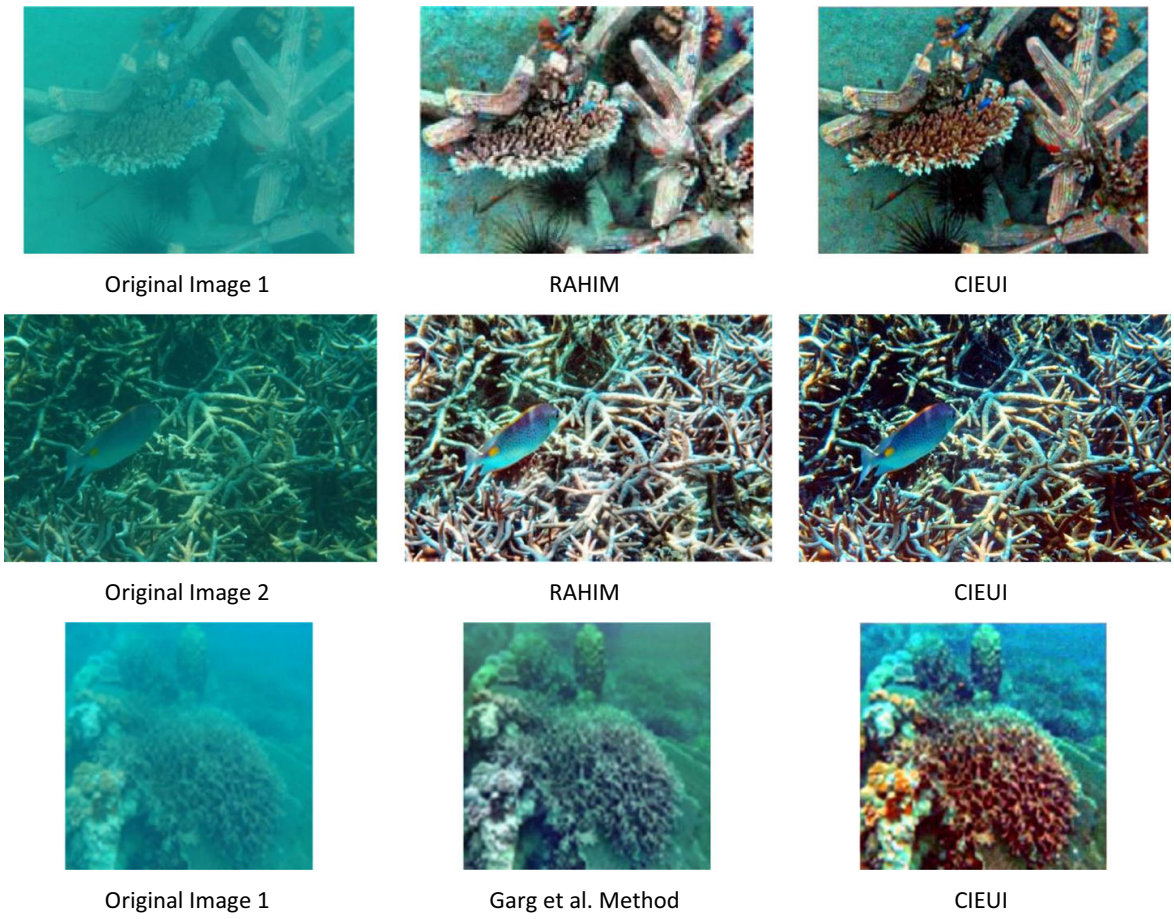


Fig. 9 Images taken from articles ([1, 16]

values due to which images have more reddish appearance.Histograms of Kwok *et al.* [31] technique images are skewed towards higher intensities (high contrast) and have non-uniform distribution due to which resultant images are very light.

Histograms of the Sethi *et al.* [45] technique cover the whole range of intensities similar to UCM [25], but have more uniform distribution than UCM [25]. The resultant images of the fish image for the Sethi *et al.* technique [45] and UCM [25] technique are almost similar and give same entropy values but contrast of Sethi *et al.* technique [45] image is slightly better than UCM [25] image which is validated by HS and UICM values of the resultant images of fish.

Some of the images of the Sethi *et al.* [45] method appear similar to that of CIEUI, e.g. scuba diver and fish image. But it can be seen from the highlighted red boxes in the resultant images of both the techniques, images of the Sethi *et al.* [45] method have a reddish tone. Histograms of blue and green channel of the scuba diver and fish images of the Sethi *et al.* [45] method show a dip for higher intensities as compared to corresponding histograms of CIEUI images, leading to the dominance of red channel in few regions of the image.

CIEUI produces images with histograms spanning the whole dynamic range. Distribution is not uniform but better than state-of-the-art techniques. Histograms of green and blue channels do not have pixels with lower intensities due to which few artifacts are present in the output images. The fish image has more artifacts due to presence of artificial light.

From Table 4, it can be seen that CIEUI improves the histogram spread value. However, it does not always outperform other techniques. It gives the best entropy value for all images except for the fish image. It gives the best UICM value for all the images except for the seal fish image. From the above analysis, it can be concluded that CIEUI always enhances the information content, color and contrast of the image and produces results that are visually more appealing and closer to reality when contrasted with state-of-the-art algorithms.

For more exhaustive comparison, The results of CIEUI have been compared with RAHIM [1] and the Garg *et al.* technique [16] which also try to tackle color and contrast issues together. A few images along with the results have been taken from the articles ([1, 16] and are shown along with the results of CIEUI in Fig. 9.

Table 5 Values of entropy, HS and UICM for images shown in Fig. 9 and Fig. 10

Image	Technique	Entropy	Histogram Spread	UICM
Image 1	Original	5.93	0.1567	0.323
	RAHIM	7.65	0.3303	4.651
	CIEUI	**7.76**	**0.3327**	**4.895**
Image 2	Original	5.86	0.1779	2.818
	RAHIM	**7.88**	0.4566	4.859
	CIEUI	7.67	**0.4945**	**4.876**
Image 3	Original	5.11	0.1827	0.465
	Garg *et al.* Method	6.78	0.1610	2.732
	CIEUI	**7.27**	**0.2454**	**3.790**
Image from TURBID dataset	Original	5.09	0.1590	0.573
	Reference	**7.13**	0.1597	**0.673**
	CIEUI	6.88	**0.1852**	**0.607**
Image from WHOI dataset	Original	5.74	0.0792	0.775
	Reference	6.37	0.1432	1.092
	CIEUI	**6.78**	**0.1785**	**1.187**

From the visual results in Fig. 9, it is clear that the images of CIEUI are visually more pleasant than the images of RAHIM [1] and the Garg *et al.* technique [16]. The contrast of the images obtained by RAHIM [1] is on the higher side and colors are not visually appealing as compared to CIEUI. On the other hand, the images obtained by the Garg *et al.* technique still have bluish tone and color cast is still left. These facts are further proven by the quantitative measures values shown in Table 5.

In Table 5, the values of quantitative measures prove that CIEUI has better information content, color and contrast as compared to RAHIM [1] and the Garg *et al.* technique. However, for image 2, RAHIM [1] has better information content but color (UICM) and contrast (HS) of CIEUI is better. However, images obtained using the Garg *et al.* technique [16] does not possess better entropy, HS and UICM values which has been proven from the visual results also. Overall, CIEUI is better than RAHIM [1] and the Garg *et al.* technique [16] in terms of color, contrast and information content of the image.

As mentioned earlier, there are two underwater datasets namely TURBID dataset [12] which is a artificially simulated database and WHOI dataset (H. Singh) which has actual underwater images with a reference image which is actually the image generated by color correction algorithm devised by WHOI scientists. Thus, for a thorough comparison, one image from each of the dataset is shown in Fig. 10. Image in the first row has been taken from TURBID dataset and Image in the second row has been taken from WHOI dataset. The value for the performance measures for the reference image, the degraded original image and CIEUI images are given in Table 5.

From the visual results shown in Fig. 10, it can be seen that the result of CIEUI for TURBID image is not as appealing as reference image but for WHOI image, result is better than the color corrected reference image which is proven by the values of performance measures as well. Result of TURBID image is having good contrast but color is not improved to that extent because the original image is an artificially simulated image. However, it is improving overall information content and contrast of the image. Thus, CIEUI produces images with improved contrast, color and information content for different types of underwater images captured in varied underwater media.

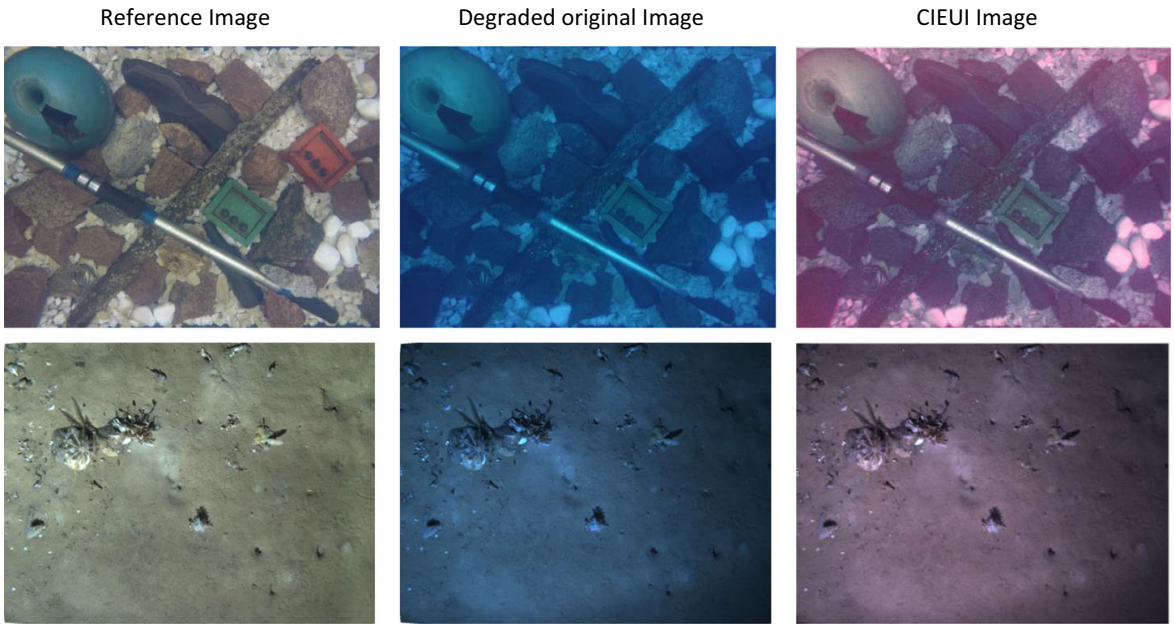


Fig. 10 First Row contains image from TURBID dataset and second row contains image from WHOI dataset

5 Conclusions

In this paper, an adaptive and optimal technique has been proposed for enhancing underwater images. The main concern of underwater images is their poor quality due to bad contrast and color which results in poor information content. Major drawbacks of state-of-the-art algorithms are (1) these methods work either on improving contrast or color or both in underwater images (2) These methods do not scale to every type of underwater image which are captured at different depth or in water bodies with varying salinity and turbidity. The key contribution of CIEUI is that it not only improves contrast and color performance but also the information content of the underwater image. It effectively avoids the artifacts (i.e. False colors and halos) caused by overstretching of the histogram. The major limitation of CIEUI is the slow execution time of MOPSO due to which it is not applicable in real time. One solution to apply the method in real time is to first get the optimized values of λ for a type of medium and then apply those values for other images captured in the same media. In this way, it can be applied in real time, and knowing the values of parameters in different type of underwater media, can be easily scaled to every type of underwater image. This method produces visually enhanced images with better color, contrast and object clarity, which is proven by quantitative measures as well. A few artifacts are still left in output images in which artificial lighting has been employed and future work will focus on removing these artifacts.

References

1. Abdul Ghani AS, Mat Isa NA (2017) Automatic System for Improving Underwater Image Contrast and Color Through Recursive Adaptive Histogram Modification. Comput Electron Agric 141(C):181–195. https://doi.org/10.1016/j.compag.2017.07.021
2. AbuNaser A, Doush IA, Mansour N, Alshattnawi S (2015) Underwater Image Enhancement Using Particle Swarm Optimization. J Intell Syst 24(1):99–115
3. Alvarez-Benitez JE, Everson RM, Fieldsend JE (2005) A MOPSO Algorithm Based Exclusively on Pareto Dominance Concepts. In: Coello Coello CA, Hernández Aguirre A, Zitzler E (eds) Evolutionary Multi-Criterion Optimization. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 459–473
4. Ancuti CO, Ancuti C, De Vleeschouwer C, Bekaert P (2018) Color Balance and Fusion for Underwater Image Enhancement. IEEE Trans Image Process 27(1):379–393. https://doi.org/10.1109/TIP.2017.2759252
5. Berkenkamp, L. (2009). Discover the Oceans: The World's Largest Ecosystem. Nomad Press
6. Buchsbaum G (1980) A spatial processor model for object colour perception. Journal of the Franklin Institute 310(1):1–26. https://doi.org/10.1016/0016-0032(80)90058-7
7. Carlevaris-Bianco N, Mohan A, Eustice RM (2010) Initial results in underwater single image dehazing. MTS/IEEE Seattle, OCEANS 2010. doi:10.1109/OCEANS.2010.5664428
8. Chambah M, Semani D, Renouf A, Courtellemont P, Rizzi A (2003) Underwater color constancy: enhancement of automatic live fish recognition. In Electronic Imaging 2004 (pp. 157–168).
9. Chao L, Wang M (2010) Removal of water scattering. ICCET 2010–2010 International Conference on Computer Engineering and Technology, Proceedings, 2, 35–39. doi:10.1109/ICCET.2010.5485339
10. Chiang JY, Chen YC (2012) Underwater image enhancement by wavelength compensation and dehazing. IEEE Trans Image Process 21(4):1756–1769. https://doi.org/10.1109/TIP.2011.2179666
11. Coello CAC, Pulido GT, Lechuga MS (2004) Handling multiple objectives with particle swarm optimization. IEEE Trans Evol Comput 8(3):256–279
12. Duarte A, Codevilla F, Gaya JDO, Botelho SSC (2016) A dataset to evaluate underwater image restoration methods. In OCEANS 2016 - Shanghai (pp. 1–6). doi:10.1109/OCEANSAP.2016.7485524
13. Galdran A, Pardo D, Picón A, Alvarez-Gila A (2015) Automatic Red-Channel underwater image restoration. J Vis Commun Image Represent 26:132–145. https://doi.org/10.1016/j.jvcir.2014.11.006
14. Gao Y, Li H, Wen S (2016) Restoration and enhancement of underwater images based on bright channel prior. Mathematical Problems in Engineering, 2016.

15. Garcia R, Nicosevici T, Cufí X (2002) On the way to solve lighting problems in underwater imaging. In OCEANS'02 MTS/IEEE. IEEE 2:1018–1024
16. Garg D, Garg NK, Kumar M (2018) Underwater image enhancement using blending of CLAHE and percentile methodologies. Multimed Tools Appl 77(20):26545–26561. https://doi.org/10.1007/s11042-018-5878-8
17. Ghani ASA, Isa NAM (2015) Underwater image quality enhancement through integrated color model with Rayleigh distribution. Appl Soft Comput 27:219–230
18. Hanmandlu M, Verma OP, Kumar NK, Kulkarni M (2009) A novel optimal fuzzy system for color image enhancement using bacterial foraging. IEEE Trans Instrum Meas 58(8):2867–2879
19. He K, Sun J, Tang X (2009) Single image haze removal using dark channel prior. In 2009 IEEE Conference on Computer Vision and Pattern Recognition (pp. 1956–1963). doi:10.1109/CVPR.2009.5206515
20. Henke B, Vahl M, Zhou Z (2013) Removing color cast of underwater images through non-constant color constancy hypothesis. In 2013 8th International Symposium on Image and Signal Processing and Analysis (ISPA) (pp. 20–24)
21. Hitam MS, Awalludin EA, Yussof WNJHW, Bachok Z (2013) Mixture contrast limited adaptive histogram equalization for underwater image enhancement. In Computer Applications Technology (ICCAT), 2013 International Conference on (pp. 1–5)
22. Hope N (n.d.) Underwater Videos and Pictures
23. Hung H-C, Chen C-L (2012) A fuzzy inference model for removing the color cast of digitally captured images under unknown illuminants. In 2012 International conference on Fuzzy Theory and Its Applications (iFUZZY2012) (pp. 192–197)
24. Iqbal K, Odetayo M, James A, Salam RA, Talib AZH (2010a) Enhancing the low quality images using unsupervised colour correction method. Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics, 1703–1709. doi:10.1109/ICSMC.2010.5642311
25. Iqbal K, Odetayo M, James A, Salam RA, Talib AZH (2010b) Enhancing the low quality images using Unsupervised Colour Correction Method. In Systems Man and Cybernetics (SMC), 2010 IEEE International Conference on (pp. 1703–1709)
26. Iqbal K, Salam RA, Osman A, Talib AZ (2007) Underwater Image Enhancement Using an Integrated Colour Model. Int J Comput Sci 34(2):239–244. https://doi.org/10.1016/S0031-3203(01)00040-1
27. James K, Russell E (1995) Particle swarm optimization. In Proceedings of 1995 IEEE International Conference on Neural Networks (pp. 1942–1948).
28. Jolla L (2015) Single underwater image enhancement using depth estimation based on blurriness Yan-Tsung Peng , Xiangyun Zhao and Pamela C. Cosman Department of Electrical and Computer Engineering , University of California , San Diego ,. International Conference on Image Processing (ICIP), 2–6.
29. Kanmani M, Narasimhan V (2017) Swarm intelligent based contrast enhancement algorithm with improved visual perception for color images. Multimed Tools Appl. https://doi.org/10.1007/s11042-017-4911-7
30. Knowles J, Corne D (1999) The pareto archived evolution strategy: A new baseline algorithm for pareto multiobjective optimisation. In Evolutionary Computation, 1999. CEC 99. Proceedings of the 1999 Congress on (Vol. 1, pp. 98–105).
31. Kwok NM, Shi HY, Ha QP, Fang G, Chen SY, Jia X (2013) Simultaneous image color correction and enhancement using particle swarm optimization. Eng Appl Artif Intell 26(10):2356–2371
32. Lam EY (2005) Combining gray world and retinex theory for automatic white balance in digital photography. In Proceedings of the Ninth International Symposium on Consumer Electronics, 2005. (ISCE 2005). (pp. 134–139). doi:10.1109/ISCE.2005.1502356
33. Lebart K, Smith C, Trucco E, Lane DM (2003) Automatic indexing of underwater survey video: algorithm and benchmarking method. IEEE J Ocean Eng 28(4):673–686
34. Li C, Guo J, Wang B, Cong R, Zhang Y, Wang J (2016) Single underwater image enhancement based on color cast removal and visibility restoration. Journal of Electronic Imaging 25(3):033012. https://doi.org/10.1117/1.JEI.25.3.033012
35. Liu Y, Liang Y, Liu S, Rosenblum DS, Zheng Y (2016) Predicting urban water quality with ubiquitous data. ArXiv Preprint ArXiv:1610.09462
36. Liu Y, Nie L, Han L, Zhang L, Rosenblum DS (2015) Action2Activity: Recognizing Complex Activities from Sensor Data. In IJCAI
37. Liu Y, Nie L, Liu L, Rosenblum DS (2016) From action to activity: Sensor-based activity recognition. Neurocomputing 181:108–115. https://doi.org/10.1016/j.neucom.2015.08.096
38. Liu Y, Zheng Y, Liang Y, Liu S, Rosenblum DS (2016) Urban Water Quality Prediction based on Multi-task Multi-view Learning (Proceeding). IJCAI 2016. Retrieved from https://www.microsoft.com/en-us/research/publication/urban-water-quality-prediction-based-multi-task-multi-view-learning-2/
39. Lu H, Li Y, Zhang Y, Chen M, Serikawa S, Kim H (2017) Underwater optical image processing: a comprehensive review. Computer Vision and Pattern Recognition:1–14. https://doi.org/10.1007/s11036-017-0863-4

40. Panetta K, Gao C, Agaian S (2016) Human-visual-system-inspired underwater image quality measures. IEEE J Ocean Eng 41(3):541–551
41. Peng YT, Cosman PC (2017) Underwater Image Restoration Based on Image Blurriness and Light Absorption. IEEE Trans Image Process 26(4):1579–1594. https://doi.org/10.1109/TIP.2017.2663846
42. Reyes-Sierra M, Coello CAC (2006) Multi-objective particle swarm optimizers: A survey of the state-of-the-art. Int J Comput Intell Res 2(3):287–308
43. Rizzi A, Gatta C, Marini D (2003) A new algorithm for unsupervised global and local color correction. Pattern Recogn Lett 24(11):1663–1677
44. Schechner YY, Averbuch Y (2007) Regularized image recovery in scattering media. IEEE Trans Pattern Anal Mach Intell 29(9):1655–1660
45. Sethi R, Sreedevi I, Verma OP, Jain V (2015) An optimal underwater image enhancement based on fuzzy gray world algorithm and Bacterial Foraging algorithm. In 2015 Fifth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG) (pp. 1–4)
46. Singh H (n.d.) \uppercase{WHOI} Color Correction Dataset.
47. Singh K, Kapoor R, Sinha SK (2015) Enhancement of low exposure images via recursive histogram equalization algorithms. Optik 126(20):2619–2625. https://doi.org/10.1016/j.ijleo.2015.06.060
48. Srinivas N, Deb K (1994) Muiltiobjective optimization using nondominated sorting in genetic algorithms. Evol Comput 2(3):221–248
49. Torres-Méndez LA, Dudek G (2005) Color correction of underwater images for aquatic robot inspection. In International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition (pp. 60–73)
50. Trelea IC (2003) The particle swarm optimization algorithm: convergence analysis and parameter selection. Inf Process Lett 85(6):317–325. https://doi.org/10.1016/S0020-0190(02)00447-7
51. Tripathi AK, Mukhopadhyay S, Dhara AK (2011) Performance metrics for image contrast. In Image Information Processing (ICIIP), 2011 International Conference on (pp. 1–4)
52. Verma OP, Kumar P, Hanmandlu M, Chhabra S (2012) High dynamic range optimal fuzzy color image enhancement using artificial ant colony system. Appl Soft Comput 12(1):394–404
53. Yang HY, Chen PY, Huang CC, Zhuang YZ, Shiau YH (2011) Low complexity underwater image enhancement based on dark channel prior. Proceedings - 2011 2nd International Conference on Innovations in Bio-Inspired Computing and Applications, IBICA 2011, 17–20. doi:10.1109/IBICA.2011.9
54. Zadeh LA (1965) Fuzzy sets. Inf Control 8(3):338–353
55. Zuiderveld K (1994) Contrast Limited Adaptive Histogram Equalization. In Graphics Gems. https://doi.org/10.1016/B978-0-12-336156-1.50061-6

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

Rajni Sethi She received her B.Tech. degree in Information Technology from Y.M.C,A, Institute of Engineering, Faridabad, India in 2009 and M.Tech degree in Information Systems from Delhi Technological University, Delhi, India in 2012. She is pursuing Ph.D. (2014) in the field of underwater image processing from Delhi Technological University, Delhi. Her research interests include image processing, color image enhancement, evolutionary computing and soft computing.

Indu Sreedevi is a Professor in the Department of Electronics and communication Engineering, Delhi Technological University, Delhi, and is currently the head of the department there. She has done Ph.D. in Computer Engineering from University of Delhi and M.Tech, Post B.Tech. from University of Kerala, Kerala. She was awarded the Best Branch Counselor award 2013 IEEE USA. Her research interests are in the field of Computer Vision, Image Processing, and wireless sensor network. She has 3 completed sponsored research projects. She published several research papers in international journals and conferences.

Microsystem Technologies
https://doi.org/10.1007/s00542-019-04545-6

TECHNICAL PAPER

Analysis of structural parameters on sensitivity of black phosphorus junctionless recessed channel MOSFET for biosensing application

Ajay Kumar[1] · Neha Gupta[2] · M. M. Tripathi[3] · Rishu Chaujar[3]

Received: 2 January 2019 / Accepted: 8 July 2019
© Springer-Verlag GmbH Germany, part of Springer Nature 2019

Abstract
This paper presents a technology computer-aided design analysis of an ultrasensitive black phosphorus junctionless recessed channel MOSFET as a biosensor. A nano cavity gap is embedded in the gate insulator region (for molecules immobilization) due to which gate capacitance changes owing to the accumulation of different molecules which reflects the deviation in threshold voltage. Higher sensitivity (1.7) is achieved for protein at low V_{ds} (0.2 V) in comparison to streptavidin (1.17) and Biotin (1.24). Further, the effect of cavity gap length and oxide thickness variation is also examined. All the results pave the way for early detection techniques of protein-related diseases such as Alzheimer's diseases, ovarian cancer and coronary artery disease with the existing complementary metal–oxide–semiconductor technology.

1 Introduction

Recent years, nanotechnology has been used to develop various kinds of biosensors for medical applications and disease diagnostics, to address its high performance. Field effect transistor (FET)-based biosensors (Bergveld 1986; Kumar et al. 2018) have gained a lot of attention due to rapid electrical detection, high scalability, low power consumption, high sensitivity, direct electrical readouts, low-cost as compared to other devices. Highly sensitive, low-cost, reliable, and quick diagnostic bio-sensing devices are vital for different biomedical applications (Vu et al. 2009). Nanotechnology-based biosensor devices are used (Azmi et al. 2014; Ahn et al. 2010; Gao et al. 2011) to overawed the problems of predictable health diagnostic methods. Constant shrinking the MOSFET at sub-20 nm regime develops opportunities for low power and high-speed applications with different gate engineering schemes with different materials. Due to very high mobility ($\sim$ 1000 cm^2 V^{-1} s^{-1}) and a layered structure (Lam et al. 2014; Kumar et al. 2018) black phosphorus (BP) is a frequently used 2D material (Cao et al. 2019). Band gap of BP can be tuned from bulk to single layer (0.3 eV to 1.5 eV) (Liu et al. 2014; Castellanos-Gomez et al. 2014; Das et al. 2014; Cao et al. 2019). Owing to these properties, BP is the utmost appropriate device material for high-performance nanoelectronics applications in comparison to other electronic device materials. To enhance the performance of the conventional grooved gate-MOSFET (Kumar et al. 2016a, b), the gate material is replaced by BP. Doping profile is uniform in the source/drain, and the substrate region to eliminate junction from the transistor and make it junctionless. JP Colinge introduced the basic properties and physics of the junctionless transistor in 2011 (Colinge et al. 2011; Sahay and Kumar 2016), and as a result, the I_{on}/I_{off} ratio, sub-threshold swing and drain induced barrier lowering (DIBL) and other properties have been enhanced due to the junctionless phenomenon.

Therefore, in this paper, for the first time, BP-JL-RC-MOSFET is introduced as a biosensor. Basically, the biosensor is used to diagnose the biological elements, attaching the binding specificity and sensitivity of biologics such as proteins, enzymes, antibodies, and nucleotides. In BP-JL-RC-MOSFET biosensor, the channel potential is

✉ Rishu Chaujar
chaujar.rishu@dtu.ac.in

Ajay Kumar
ajaykumar@dtu.ac.in

Neha Gupta
gupta.neha@niecdelhi.ac.in

M. M. Tripathi
mmtripathi@dce.ac.in

[1] ECE Department, Jaypee Institute of Information Technology, Noida, India

[2] Department of Applied Science and Humanity, ADGITM, New Delhi, India

[3] Delhi Technological University, Delhi, India

Published online: 13 July 2019

influenced by the electrical properties of biomolecules. In conventional MOSFET, when gate voltage surpasses the threshold voltage (V_{TH}), then the current flow in the channel, and it can be altered by the dielectric constant of biomolecule near the gate electrode and hence, threshold voltage changes due to change in gate capacitance. In BP-JL-RC-MOSFET, in the gate insulator region, a nano-cavity is embedded for immobilizing the biomolecules (Lai et al. 2012). The electrical characteristics are altered due to absorption of biomolecules and due to which the sensitivity of sensor changes. The biomolecules which are absorbed in the Nano-cavity gap changes the dielectric constant (k) of the gap which alters the V_{TH} of the device which then serves as a key parameter to estimate the sensitivity of marker.

Moreover, steric hindrance issue has been considered since steric hindrance is the slowing of chemical reactions due to steric bulk. It is usually demonstrated in intermolecular reactions, whereas discussion of steric effects often focuses on intermolecular interactions (Mahshid et al. 2015; Schuck et al. 1999). Steric hindrance is often exploited to control selectivity, such as slowing unwanted side-reactions. As our analysis is based on simulation (which is calibrated with the previously published experimental work), it has been assumed that the steric hindrance will occur but by choosing small sized molecules, this effect can be eliminated as there is virtually no upper limit on the size of the analyte (Schuck et al. 1999). Moreover, since we are proposing a nano biosensor, small sized particles are preferred. Also high surface coverage may lead to steric hindrance of analyte binding to neighboring surface sites that is why, the nano cavity gap is not large and sufficient for biomolecules immobilizing.

2 Device structure and its parameters

The device design for BP-JL-RC-MOSFET considered in this work is shown in Fig. 1. The difference between nano cavity-gap implanted BP-JL-RC-MOSFET and conventional recessed channel MOSFET is that it involves the nano cavity-gap at the edge of the gate dielectric where the biomolecules are accumulated. Here, L_{gap} is the Nano-gap cavity length (8 nm), t_{ox} (oxide thickness) is 2 nm, L_G (gate length) 20 nm, and width of the device is 200 nm. The doping of source/drain region are same to create a junctionless device and doped with an n-type impurity of 5×10^{17} $cm^{-3,}$ and the substrate is also doped with an n-type impurity of 5×10^{17} cm^{-3}. Negative Junction Depth (NJD) is 10 nm, V_{gs} is 1.5 V, and V_{ds} is 0.2 V. The absorption of various biomolecules is modeled by introducing insulator with dielectric constant same as that of a particular biomolecule in the Nano-gap cavity.

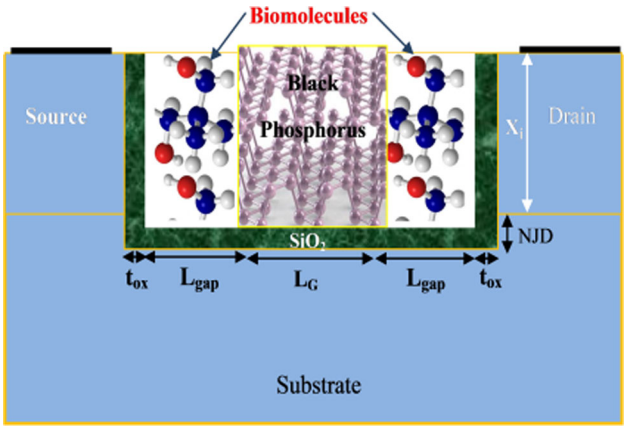


Fig. 1 The device design for BP-JL-RC-MOSFET

3 Simulation methodology and calibration

In this work, the ATLAS device simulator has been used (Silvaco 2011). In this simulation, FLDMOB (Parallel Electric Field Dependence) model is included for mobility of carriers under the influence of changing electric field. KLASRH (concentration-dependent Klaassen Shockley–Read–Hall Recombination) model is considered for carrier generation–recombination as it can include a concentration-dependent lifetime of carriers (Silvaco 2011). In the simulation, the gate voltage is varied from 0 to 1.5 V while drain bias is kept constant at 0.2 V.

The following equations (Silvaco 2011) and models have been used in the simulation process of the device.

1. Poisson's equation

$$div(\varepsilon \nabla \varphi) = -\rho \tag{1}$$

where ε, ρ and φ represents permittivity, space charge density and potential respectively.

2. Current continuity equation

For Holes,

$$\frac{\partial p}{\partial t} = \frac{1}{q} div \vec{J}_p + G_p - R_p \tag{2}$$

For Electrons,

$$\frac{\partial n}{\partial t} = \frac{1}{q} div \vec{J}_n + G_n - R_n \tag{3}$$

where $\vec{J}_p$ and $\vec{J}_n$ are holes and electrons current densities; the generation rate for holes (p) and electrons (n) repre-

sented by G_p and G_n respectively; R_p is the holes recombination rate and R_n is electrons recombination rate and 'q' represents for charge.

4 Recombination

4.1 SRH (Shockley–Read–Hall)

$$R_{SRH} = \frac{pn - n_{ie}^2}{\tau_p[n + n_{ie}e^{(-(E_{TRAP}/kT_L))}] + \tau_n[p + n_{ie}e^{(-(E_{TRAP}/kT_L))}]} \quad (4)$$

where change in the intrinsic Fermi level and the trap energy level represented by E_{TRAP}; τ_n and τ_p, are the electron and hole lifetimes.

4.2 Auger recombination

$$R_{Auger} = AUGN\left(pn^2 - nn_{ie}^2\right) + AUGP\left(np^2 - pn_{ie}^2\right) \quad (5)$$

AUGP and AUGN are user-defined where AUGP = 1.8 × 10^{-31} cm^6/s and AUGN = 8.3 × 10^{-32} cm^6/s.

5 Parallel electric-field-dependent mobility

$$\mu(E) = \mu_0\left[\frac{1}{1 + (\mu_0 E/\upsilon_{sat})^{\beta}}\right]^{1/\beta} \quad (6)$$

Simulation models have been validated by both the experimental statistics (Xiao-Hua et al. 2006) and modeled statistics (Sreelal et al. 2002). Figure 2a reflected the calibration of experimental data (Xiao-Hua et al. 2006) and simulated data and plotted for 140 nm gate length device. Hence, we have approximated the corner effect (Chaujar et al. 2008) and find out that the error ($\Delta\varepsilon/L_{eff}$) < 0.6%. This error has been neglected in this analysis (Fig. 2a). Moreover, models calibration has been performed using modeled data (Sreelal et al. 2002) and it is observed in Fig. 2b that the modeled and simulated results are well calibrated, thus showing the validity of simulation models. Furthermore, to show the device fabrication feasibility, the process flow of proposed device fabrication is shown in Fig. 3. When the device is fabricated, biomolecules used in biosensor must pass through both pre and post processing steps to ensure high purity and thus accurate sensitivity as the pre- and post- processing treatments of biomolecules are essential during the fabrication of biosensors.

6 Results and discussion

6.1 Electrical detection of biomolecules

It is observed from Fig. 4a, when the biomolecules are immobilized (k > 1) in the Nano-gap cavity, ON-current rises significantly in comparison to the absence of biomolecule (k = 1) owing to increased dielectric constant which increases the gate control. Therefore, the switching ratio (I_{ON}/I_{OFF}) also increases with a dielectric constant (streptavidin, biotin, and protein respectively) as compared to air (unfilled). It is also superior than conventional MOSFET as reflected in Table 1 and hence, can be used as a key parameter to detect/sense the specific biomolecule in the BP-JL-RC device as shown in Fig. 4b due to the amalgamation of black phosphorus and junctionless phenomenon. Further, the sensitivity (S) of the device has been calculate and defined as given in Eq. (7)

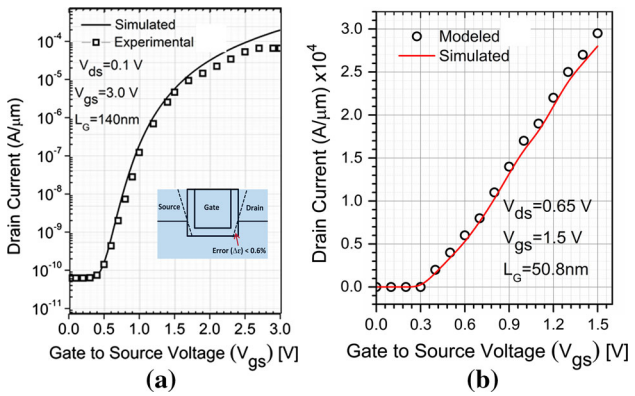


Fig. 2 **a** Calibration of $I_{ds} - V_{gs}$ characteristics of grooved gate MOSFET: **a** L_G = 140 nm by experimental (Xiao-Hua et al. 2006) and simulation data. **b** L_G = 50.8 nm by modelled (Sreelal et al. 2002) and simulation data

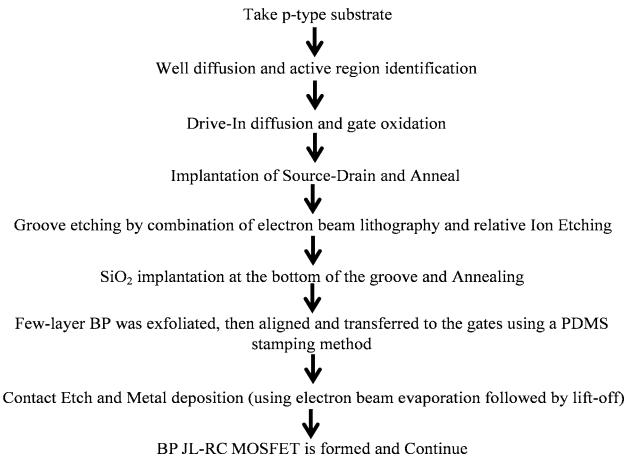


Fig. 3 Fabrication process flow of BP-JL-RC-MOSFET

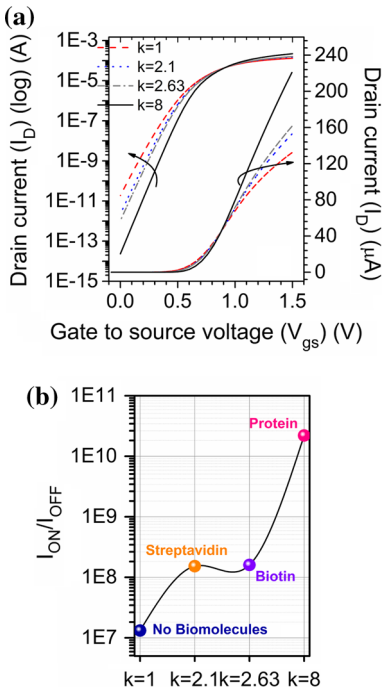


Fig. 4 **a** Transfer characteristics, and **b** switching ratio (I_{ON}/I_{OFF}) of BP-JL-RC-MOSFET in the presence of biomolecules

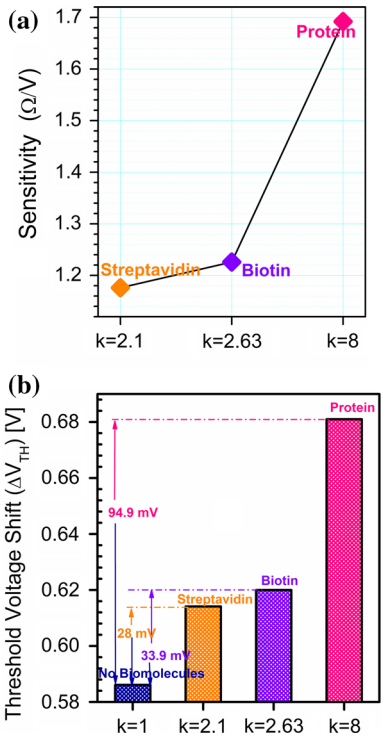


Fig. 5 **a** Impact on sensitivity and **b** threshold voltage shifts for different biomolecules in BP-JL-RC-MOSFET

Table 1 Device performance comparison of conventional MOSFET and BP-JL-RC-MOSFET

Parameter	Sensitivity	ΔV_{TH}	I_{on}/I_{off}
Conventional MOSFET			
k = 2.1	0.9	2 mV	9365
k = 2.63	1.009	5 mV	9832
k = 8	1.16	9 mV	10,003
BP-JL-RC MOSFET			
k = 2.1	1.176	28 mV	1.52E8
k = 2.63	1.226	33.9 mV	1.6E8
k = 8	1.692	94.9 mV	2.2E10

$$S = \frac{I_{ON}(k > 1)}{I_{ON}(k = 1)} \tag{7}$$

Figure 5a shows the sensitivity of BP-JL-RC-MOSFET. It is evaluated that the sensitivity for streptavidin, biotin and protein biomolecules are 1.17, 1.24 and 1.7 respectively and also better than conventional MOSFET (shown in Table 1). The enhancement in sensitivity is due to the increment in I_{ON} as well as switching ratio (shown in Fig. 4a, b). From the figure, it is also evident that the sensitivity is comparatively higher for the protein biomolecules owing to a high dielectric constant at lower drain bias (0.2 V at constant V_{gs} = 1.5 V) as compared to streptavidin and biotin. Thus, BP-JL-RC device is also applicable for low power ultra-sensing applications in coronary artery disease, ovarian cancer, and Alzheimer's disease diagnosis.

The threshold voltage is also used for detection purpose as a bio-sensing parameter, and it is directly related to the dielectric constant. With the increase of particular dielectric constant of a biomolecule, threshold voltage enhances, and the same results are observed for BP-JL-RC-MOSFET as shown in Fig. 5b. High threshold voltage shift (ΔV_{TH}) of 28 mV, 33.9 mV, and 94.9 mV are observed in BP-JL-RC device when streptavidin, biotin, and protein respectively are immobilized in the Nano-gap cavity in comparison to unfilled Nanogap cavity. Higher ΔV_{TH} is observed in proposed biosensor as compared to conventional MOSFET biosensor as reflected in Table 1.

6.2 Impact on stray capacitances

The presence of biomolecule can be easily observed by the change in stray capacitance (Kumar et al. 2017) because the stray or parasitic capacitances of a FET gets changed by altering the respective dielectric constant. From Fig. 6a it is observed that the gate capacitance (C_{gg}) is altered (decreased) by altering (increasing) the corresponding dielectric constant of streptavidin, biotin, and protein in compared to air. Similarly, gate to source capacitance (C_{gs}) and the gate to drain capacitance (C_{gd}) are altered in the presence of different biomolecules as shown in Fig. 6b, c.

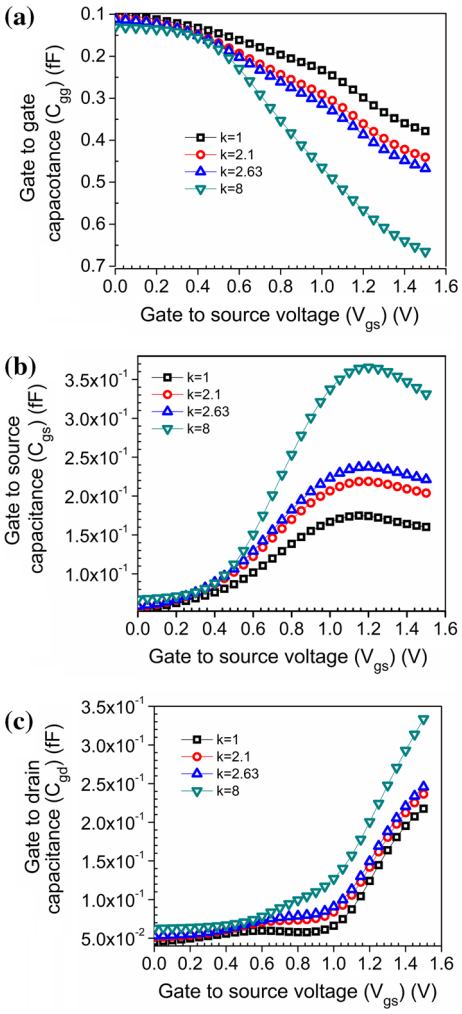


Fig. 6 Effect on parasitic capacitances **a** C_{gg} **b** C_{gs} **c** C_{gd}; in the presence of biomolecules in BP-JL-RC-MOSFET

When the protein molecules are immobilized, the change in the capacitances is more prominent for protein while less prominent for biotin and streptavidin molecules as compared to air as evident from Fig. 6a–c.

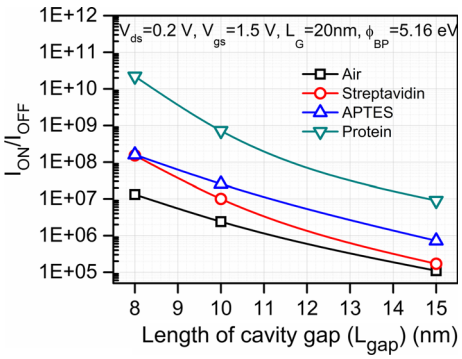


Fig. 7 The impact of length of cavity gap variation on I_{ON}/I_{OFF}, for different biomolecules in BP-JL-RC-MOSFET

Deflection in parasitic capacitance due to biomolecule also deflected the characteristics of the MOSFET. Thus, with the change in parasitic capacitances, the presence of biomolecule can easily be detected.

6.3 Effect of nano-gap cavity length

Moreover, the effect of cavity length (L_{gap}) has also been observed in this work with an aim to investigate the sensitivity at different L_{gap} of the biosensor. Figure 7a shows the impact of L_{gap} variation on I_{ON}, and it is observed that I_{ON} increases for streptavidin, APTES, and protein as compared to air for a particular L_{gap} (8 nm, 10 nm, and 15 nm) owing to reduction in parasitic capacitances (as shown in Fig. 6) but I_{ON} reduces when L_{gap} increases from 8 nm to 15 nm. Correspondingly, the switching ratio is increased for streptavidin, APTES, and protein respectively as compared to air for a particular L_{gap} reflected in Fig. 7b due to higher I_{ON} while it reduces with increased L_{gap} from 8 to 15 nm. Figure 8 shows the change (reduces) in threshold voltage due to change in L_{gap} (from 8 nm to 15 nm) for air, streptavidin, APTES, and protein respectively. Thus, a prominent change in threshold voltage (due to the change in dielectric constant) is used as a sensing parameter for detection of protein molecules. Moreover, the sensitivity of BP-JL-RC-MOSFET is also observed with the change in L_{gap}, and it is found that sensitivity is enhanced significantly with the increase in L_{gap} as shown in Fig. 9. From the figure, it is also observed that the sensitivity increases from streptavidin to APTES and APTES to protein (for each L_{gap}). Thus, with L_{gap} = 15 nm, Nano-gap embedded.

BP-JL-RC-MOSFET shows improved sensitivity for protein biomolecules as compared to other biomolecules. Thus, 15 nm cavity-gap length of BP-JL-RC-MOSFET show optimum performance and hence can be used as a low power, high sensitive biosensor to diagnose various associated diseases.

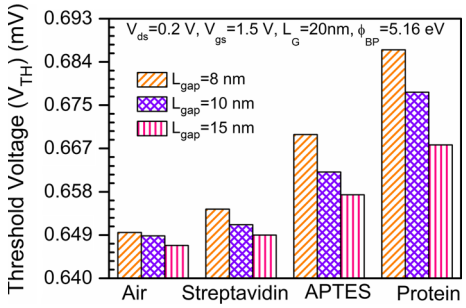


Fig. 8 The impact of length of cavity gap variation on threshold voltage, for different biomolecules in BP-JL-RC-MOSFET

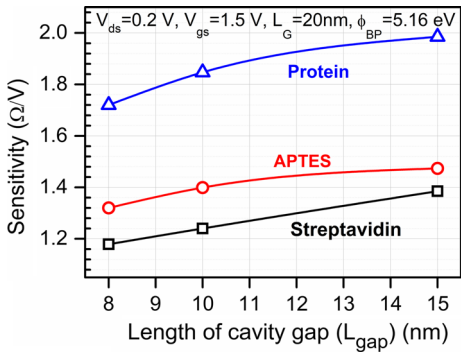


Fig. 9 The impact of length of cavity gap variation on sensitivity, for different biomolecules in TGRC-MOSFET

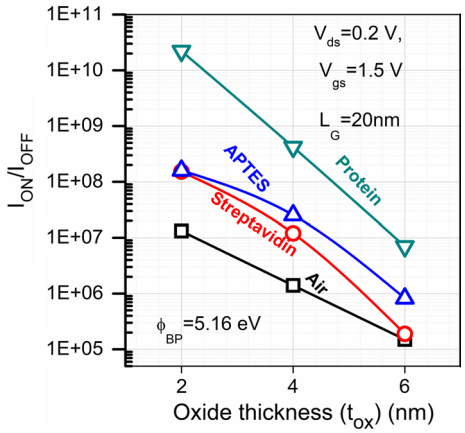


Fig. 10 The impact of oxide-thickness variation on I_{ON}/I_{OFF}, for different biomolecules in BP-JL-RC-MOSFET

6.4 Impact of oxide-thickness: limit of detection

In this section, the impact of oxide-thickness (t_{ox}) variation in order to determine the limit of detection has observed in terms of on-current, switching ratio, threshold voltage and

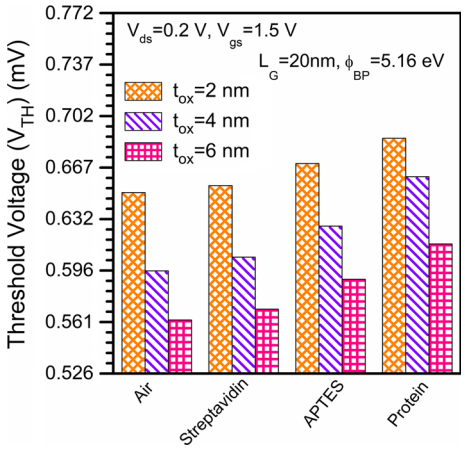


Fig. 11 The impact of oxide-thickness variation on Threshold voltage, for different biomolecules in BP-JL-RC-MOSFET

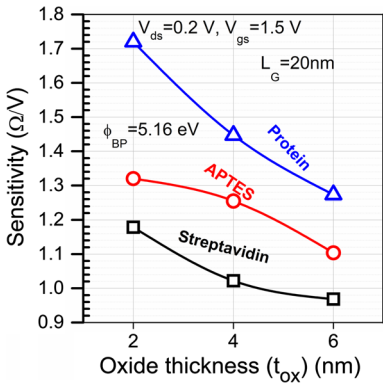


Fig. 12 The impact of oxide-thickness variation on Sensitivity, for different biomolecules in BP-JL-RC-MOSFET

sensitivity in BP-JL-RC-MOSFET. In this section, cavity gap is fixed at 8 nm because; the device gives high performance at 8 nm L_{gap} as a biosensor.

Figure 10 shows the impact of t_{ox} variation on switching ratio, and it is observed that switching ratio increases for streptavidin, APTES, and protein as compared to air for a particular t_{ox} (2 nm, 4 nm, and 6 nm) but switching ratio reduces when t_{ox} increases from 2 nm to 6 nm due to higher t_{ox}, gate capacitance (stray capacitance) will be high thus switching ratio reduced.

Figure 11 shows the change (reduced) in threshold voltage due to change in t_{ox} (from 2 nm to 6 nm) for air, streptavidin, APTES, and protein respectively. Thus, more prominent change in threshold voltage (due to change in dielectric constant) is used as sensing parameter for detection of biomolecules. Moreover, the sensitivity of the proposed device is also observed for different t_{ox}, and it is observed that the sensitivity is reduced with the increase in t_{ox} as shown in Fig. 12 and maximum at 2 nm t_{ox} due to higher I_{ON}. Thus, 2 nm oxide-thickness of BP-JL-RC device show optimum performance and hence can be used as a low power, higher sensitive biosensor. From the results, it is also concluded that there is a detection limit of up to 4 nm oxide thickness and beyond this limit the device is not suitable for bio-sensing applications due to higher reduction in I_{ON} and sensitivity.

7 Conclusion

In this work, BP-JL-RC-MOSFET has been proposed as a biosensor for electrical detection of biomolecules. The device performance (as a biosensor) has been investigated in terms of sensitivity, V_{TH} shift, I_{ON}/I_{OFF} ratio, and drain current. Due to higher sensitivity at low drain bias, all the parameters have been investigated at 0.2 V drain voltage for different biomolecules and also in the absence of molecules (air). Further, the efficacy of cavity gap length

and oxide thickness variation has also been observed on I_{ON}, I_{ON}/I_{OFF}, V_{TH}, and sensitivity. Thus, BP-JL-RC-MOSFET proves to be a promising device for biosensing applications. Owing to high sensitivity, it is suitable for early detection of deadly diseases and could improve the health of at-risk populations worldwide.

Acknowledgements The authors are thankful to Jaypee Institute of Information Technology and Delhi Technological University and for supporting this work.

References

Ahn J-H, Choi S-J, Han J-W, Park TJ, Lee SY, Choi Y-K (2010) Double-gate nanowire field effect transistor for a biosensor. Nano Lett 10(8):2934–2938

Azmi MM, Tehrani Z, Lewis R, Walker K-A, Jones D, Daniels D, Doak S, Guy O (2014) Highly sensitive covalently functionalised integrated silicon nanowire biosensor devices for detection of cancer risk biomarker. Biosens Bioelectron 52:216–224

Bergveld P (1986) The development and application of FET-based biosensors. Biosensors 2(1):15–33

Cao MS, Shu JC, Wang XX, Wang X, Zhang M, Yang HJ, Fang XY, Yuan J (2019a) "Electronic structure and electromagnetic properties for 2D electromagnetic functional materials in gigahertz frequency. Annalen der Physik 531(4):1800390

Cao MS, Wang XX, Zhang M, Shu JC, Cao WQ, Yang HJ, Fang XY, Yuan J (2019b) Electromagnetic response and energy conversion for functions and devices in low-dimensional materials. Adv Funct Mater 29(1807398):1–54

Castellanos-Gomez A, Vicarelli L, Prada E, Island JO, Narasimha-Acharya K, Blanter SI, Groenendijk DJ, Buscema M, Steele GA, Alvarez J (2014) Isolation and characterization of few-layer black phosphorus. 2D Materials 1(2):025001

Chaujar R, Kaur R, Saxena M, Gupta M, Gupta R (2008) Laterally amalgamated DUal material GAte concave (L-DUMGAC) MOSFET for ULSI. Microelectron Eng 85(3):566–576

Colinge J, Lee C, Akhavan ND, Yan R, Ferain I, Razavi P, Kranti A, Yu R (2011) Junctionless transistors: physics and properties. In: Semiconductor-on-insulator materials for nanoelectronics applications. Springer, pp 187–200

Das S, Zhang W, Demarteau M, Hoffmann A, Dubey M, Roelofs A (2014) Tunable transport gap in phosphorene. Nano Lett 14(10):5733–5739

Gao A, Lu N, Dai P, Li T, Pei H, Gao X, Gong Y, Wang Y, Fan C (2011) Silicon-nanowire-based CMOS-compatible field-effect transistor nanosensors for ultrasensitive electrical detection of nucleic acids. Nano Lett 11(9):3974–3978

Kumar A, Gupta N, Chaujar R (2016a) Analysis of novel transparent gate recessed channel (TGRC) MOSFET for improved analog behaviour. Microsyst Technol 22(11):2665–2671

Kumar A, Gupta N, Chaujar R (2016b) TCAD RF performance investigation of transparent gate recessed channel MOSFET. Microelectron J 49:36–42

Kumar A, Tripathi M, Chaujar R (2017) Investigation of parasitic capacitances of In_2O_5Sn gate electrode recessed channel MOSFET for ULSI switching applications. Microsyst Technol 23(12):5867–5874

Kumar A, Tripathi MM, Chaujar R (2018a) Ultralow-power dielectric-modulated nanogap-embedded sub-20-nm TGRC-MOSFET for biosensing applications. J Comput Electron 17(4):1807–1815

Kumar A, Tripathi M, Chaujar R (2018b) Comprehensive analysis of sub-20 nm black phosphorus based junctionless-recessed channel MOSFET for analog/RF applications. Superlattices Microstruct 116:171–180

Lai WA, Lin CH, Yang YS, Lu MS (2012) Ultrasensitive and label-free detection of pathogenic avian influenza DNA by using CMOS impedimetric sensors. Biosens Bioelectron 35(1):456–460

Lam K-T, Dong Z, Guo J (2014) Performance limits projection of black phosphorous field-effect transistors. IEEE Electron Device Lett 35(9):963–965

Liu H, Neal AT, Zhu Z, Luo Z, Xu X, Tománek D, Peide DY (2014) Phosphorene: an unexplored 2D semiconductor with a high hole mobility. ACS Nano 8:4033–4041

Mahshid SS, Camiré SB, Ricci F, Vallée-Bélisle A (2015) A highly selective electrochemical DNA-based sensor that employs steric hindrance effects to detect proteins directly in whole blood. J Am Chem Soc 137(50):15596–15599

Sahay S, Kumar MJ (2016) Realizing efficient volume depletion in SOI junctionless FETs. IEEE J Electron Devices Soc 4(3):110–115

Schuck P, Boyd LF, Andersen PS (1999) Measuring protein interactions by optical biosensors. Curr Protoc protein Sci 17(1):20.2.1–20.2.22

Silvaco I (2011) ATLAS user's manual. Santa Clara, CA, Ver, vol 5

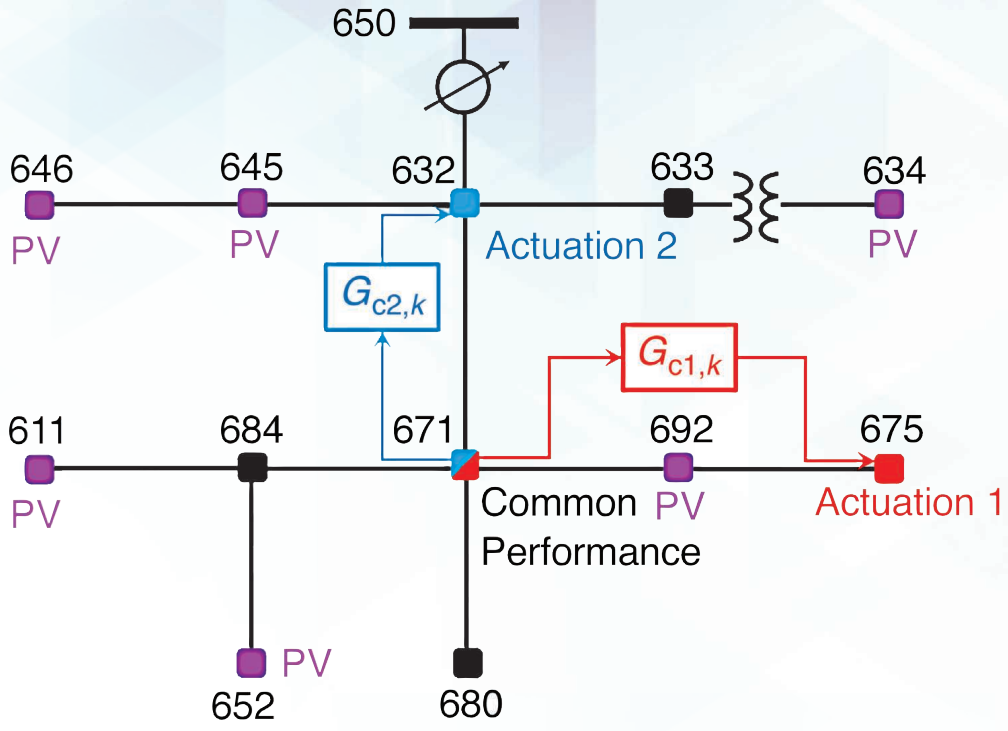
Sreelal S, Lau C, Samudra G (2002) Parasitic capacitance characteristics of deep submicrometre grooved gate MOSFETs. Semicond Sci Technol 17(3):179

Vu XT, Eschermann JF, Stockmann R, GhoshMoulick R, Offenhäusser A, Ingebrandt S (2009) "Top-down processed silicon nanowire transistor arrays for biosensing. Phys Status Solidi (a) 206(3):426–434

Xiao-Hua M, Yue H, Bao-Gang S, Hai-Xia G, Hong-Xia R, Jin-Cheng Z, Jin-Feng Z, Xiao-Ju Z, Wei-Dong Z (2006) Fabrication and characterization of groove-gate MOSFETs based on a self-aligned CMOS process. Chin Phys 15(1):195

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

Phasor-Based Adaptive Control of a Test-Feeder Distribution Network



APPLICATION OF RETROSPECTIVE COST ADAPTIVE CONTROL TO THE IEEE 13-NODE TEST FEEDER

SYED ASEEM UL ISLAM, ELIZABETH L. RATNAM, ANKIT GOEL, and DENNIS S. BERNSTEIN

Control in the power distribution system is needed to match the generation of electricity (that is, supply) with the demand for power (that is, load) at all times while maintaining a high level of power quality [1] (see "Summary"). The complexity of controlling the electric distribution system increases when distributed intermittent energy sources, such as solar photovoltaics (PVs), displace large-scale dispatchable generation [2]–[10]. If the supply from distributed resources exceeds the time-varying demand across a feeder, then the feeder-level voltage will increase. In extreme cases, large voltage increases will trigger the disconnection of distributed generation to protect electrical equipment and ensure customer safety. Consequently, coordinating renewable energy sources that are distributed throughout the distribution system—to match real-time demand across an entire power network—brings the additional challenge of maintaining a high level of supply reliability.

Digital Object Identifier 10.1109/MCS.2019.2913611
Date of publication: 18 July 2019

1066-033X/19©2019IEEE

To accommodate both traditional and alternative energy sources in the electrical distribution system, it is necessary to improve the power quality across a distribution feeder. More specifically, both time-varying generation and demand can be viewed as disturbances [11]–[13] since they cause undesirable fluctuations and offsets in the voltage magnitudes and angles in a three-phase power network. To improve power quality and thus prevent damage to electrical equipment on the customer and supply sides, voltage magnitudes must be corrected. Furthermore, to ensure that thermal constraints are not exceeded and reverse power flows are controlled, voltage angles must be corrected. The goal of this article is to develop a methodology for correcting voltage magnitudes and angles in three-phase power, thus improving power quality.

Model-based methods for the optimization and control of power systems are given in [14] and [15]. In contrast to these methods, this article develops a methodology that reflects conditions that often occur when a controller is developed for a physical system. In particular, a representative simulation of the system that captures key details may be available; however, this simulation may be erroneous due to unmodeled dynamics and unknown parameters. In addition, the representative simulation of the system may be available only in the form of a simulation code, the details of which are not accessible to the user. This can occur when the code uses extensive lookup tables, involves complex physics, or is proprietary. In these cases, the simulation is effectively an opaque executable simulation.

Motivated by these constraints, an approach whereby an adaptive controller is tuned based on the nominal simulation is considered in this article. However, this approach does not attempt to identify the system, which may be a nontrivial task due to its complexity. The simulation can be run multiple times to tune the weightings of the adaptive controller, which is referred to as *nominal simulation tuning* (NST). The weightings of the adaptive controller obtained under NST are then fixed, and the robustness and performance of the adaptive controller with the fixed weightings are evaluated through subsequent simulations based on modifications of the nominal simulation, which is referred to as *perturbed simulation testing* (*PST*). These modifications are chosen to reflect off-nominal conditions of the model, commands, disturbances, and noise that can potentially occur in real-world operation. These off-nominal conditions are merely representative, however, and not intended to be exhaustive. Therefore, PST provides confidence in the ability of retrospective cost adaptive control (RCAC) to adapt to the actual real-world system. Without knowledge of a model set that contains the actual real-world system, however, this approach does not guarantee success.

If the performance and robustness of the fixed-weighting adaptive controller are found to be acceptable for the perturbed simulations, then the adaptive controller is considered viable for the real-world system. The success of this approach depends on the extent to which the nominal and off-nominal simulations capture the key features of the real-world system. Furthermore, success in real-world operation depends on the ability of the adaptive controller to accommodate features of the real-world system that were not captured by either the nominal or perturbed simulations. The difference between this methodology and classical control synthesis is the limited access that the engineer has to plant modeling details, relying, to a large extent, on the ability of the adaptive controller to adapt to discrepancies among the nominal simulations, the perturbed simulations, and the real-world system.

This article applies RCAC to the IEEE 13-node test feeder (IEEE13NTF), one of several internationally recognized and widely studied distribution system models [16]. Estimation, optimization, fault analysis, and control studies using IEEE13NTF are given in [17]–[30]. This three-phase, four-wire IEEE13NTF model captures the unbalanced nature of the distribution network with single-phase laterals and single-phase loads. Numerical testing of this model shows that convergence after a load disturbance is typically achieved in fewer than approximately 10 steps. Because of the nonlinearity and complexity of the three-phase power flows in this model, this article views IEEE13NTF as an opaque executable simulation, with no access to the underlying code and equations. The version of IEEE13NTF used in this article is provided by the OPAL-RT RT-Lab ePHASORSIM package [31]. This version of IEEE13NTF is proprietary, and the underlying code is inaccessible.

Summary

An electrical grid consists of multiple interconnected components that provide and use energy. A key concern for these systems is the ability to maintain the desired voltage magnitudes and angles throughout the grid in the presence of time-dependent disturbances in the form of renewable energy sources. This is a decentralized feedback control problem, where modeling information is limited by the changing topology and parameters of the grid. We consider command-following and disturbance-rejection problems for a standard grid model, the IEEE 13-node test-feeder model. The model is quasi-static due to the assumption of fast transient response. Retrospective cost adaptive control (RCAC) is applied in both centralized and decentralized control architectures, where each multiple input, multiple output controller has six inputs and six outputs, due to the need to regulate the magnitude and angle of three-phase power at each power node. The performance of RCAC is evaluated under conditions of extremely limited model information.

The control objective for IEEE13NTF is to achieve setpoint command-following for three-phase voltage magnitude and angle in the presence of load and generation disturbances. To meet the control objective, we consider both single-controller and decentralized dual-controller architectures with various choices of the performance, actuation, and disturbance nodes. The performance and actuation nodes are three phase, and the disturbance nodes are either single or three phase. The performance variables at each performance node are the command-following errors for the three voltage magnitudes and three voltage angles.

The approach of this article is based on RCAC, which is a data-driven, adaptive control method for output-feedback stabilization, command-following, and disturbance rejection [32], [33]. For single-input, single-output (SISO) linear plants, RCAC uses minimal modeling information, namely, the sign of the leading numerator coefficient, relative degree, and nonminimum-phase (NMP) zeros.

The objective of this article is to apply NST and PST based on RCAC to the IEEE13NTF model. As explained in [32], the first step in applying RCAC to the IEEE13NTF model is to construct the target model G_f for the intercalated transfer function. Within the context of linear systems, this modeling information contains knowledge about the leading signs, relative degree, and NMP zeros. As shown in [32], [34], and [35], RCAC is also effective on nonlinear systems with input and feedback nonlinearities. For the IEEE13NTF model, the required modeling information is obtained from a single simulation scenario. The relevant information consists of the sign of the first impulse-response matrix of the three-phase voltage magnitudes and angles, in response to an impulse in the active and reactive power. Specifically, a single 6×6 impulse-response matrix H_1 is used to construct a finite-impulse-response (FIR) target model G_f. The dependence of the closed-loop performance on H_1 is examined by determining those components of H_1 that can be removed without adverse effect on the performance of the nominal simulation. This investigation leads to a sparse approximation, $\hat{H}_1$, of H_1. In fact, $\hat{H}_1$ is diagonal for a specific pairing between the load actuation input and the voltage magnitude and angle. This procedure, along with selection of the controller order and structure and additional weightings, yields the fixed controller tuning for use in PST.

Next, the nominal simulation is perturbed in various ways that reflect variations in how RCAC may be used in the real-world system. Specifically, the controller structure and weightings obtained from NST are first applied in a centralized architecture, where a single multiple-input, multiple-output (MIMO) controller is applied between selected actuation and performance nodes. Next, a pair of MIMO subcontrollers is applied in a decentralized architecture, where each RCAC subcontroller operates independently of the other subcontroller, without information exchange or hierarchical coordination. In particular, we apply two RCAC subcontrollers in a decentralized architecture to determine the possibility of adverse interactions between controllers due to the lack of communication. In both scenarios, the only modeling information used by RCAC is the sparse matrix $\hat{H}_1$ used to define G_f.

In both the centralized and decentralized scenarios, the control input at each actuation node is the active and reactive power extracted from each of the three phases. The feedback controller that connects each performance node to a corresponding actuation node is, therefore, a six-input, six-output MIMO controller. Each 6×6 controller is fully populated and thus consists of 36 SISO controllers, each of which is a linear, strictly proper, discrete-time dynamic compensator. Since RCAC adapts the coefficients of the controller, it is a linear time-varying controller; however, after the controller coefficients converge, it is a linear time-invariant controller. For decentralized control, each subcontroller is of the same type. As a test of the decentralized, fixed-structure subcontroller, we consider the case of multivariable actuator saturation, which corresponds to a physical limitation on the maximum power flow through a four-quadrant operation inverter. The level and form of the saturation are unknown to the subcontrollers.

Finally, we consider decentralized control in the case in

TABLE 1 **A summary of simulations used for nominal simulation tuning (NST) and perturbed simulation testing (PST). The first simulation represents the case chosen for NST. Each subsequent simulation represents a perturbation of the nominal simulation in terms of the disturbance signals; actuator saturation levels; and choice of disturbance, actuation, and performance nodes.**

Simulation	Centralized (C) or Decentralized (D)	Disturbance Nodes	Actuation Nodes	Performance Nodes	Sensor Noise	Saturation
NST	C	611, 634	675	671	Yes	No
PST 1	C	611, 634	632	633	Yes	No
PST 2	D	611, 634	632, 675	633, 671	Yes	No
PST 3	D	611, 634	632, 675	671	Yes	Yes
PST 4	D	Photovoltaics at 611, 634, 645, 646, 652, 692	632, 675	671	Yes	No

which the IEEE13NTF model includes significant PV penetration. PV generation data from Pecan Street [36], an online repository of data collected from real consumers and producers of electricity, are used. The data from various consumers/producers are aggregated to simulate loads on nodes in the IEEE13NTF model. This method is used to simulate 70.61% PV penetration in the distribution system. A summary of the numerical simulations is presented in Table 1. The nominal simulation is run multiple times to tune the RCAC weightings, select the controller structure, and construct the sparse G_f. PSTs 1–4 are run once by using the controller from NST with the modifications to the nominal simulation presented in Table 1.

The contribution of this article is the application of RCAC to IEEE13NTF, for which the only modeling information used is a sparse approximation of the first impulse-response coefficient. An additional contribution of this article is the use of a controller structure that simplifies the controller structure used in [32]. Specifically, the approach of [32] specifies the order of the controller in MIMO input–output model form; otherwise, the controller is unstructured in terms of the allowable locations of its poles and zeros. In contrast, this article employs a variation of [32], called *fixed-structure RCAC*, whereby the user can specify the internal structure of each SISO entry of the 6×6 MIMO controller. Since each SISO entry of the 6×6 MIMO controller is specified as either a proportional-integral (PI) controller or a second-order infinite-impulse-response (IIR) controller, this extension reduces the complexity of feedback control of IEEE13NTF.

IEEE13NTF MODEL

Consider the nonlinear, time-varying, discrete-time plant

$$x(k+1) = f(x(k), u(k), d(k), k), \tag{1}$$

$$y_0(k) = g(x(k), u(k), d(k), k), \tag{2}$$

$$y_n(k) = y_0(k) + v(k), \tag{3}$$

$$z(k) \triangleq r(k) - y_n(k), \tag{4}$$

where $k \geq 0$ is the step, $x(k) \in \mathbb{R}^n$ is the state, $u(k) \in \mathbb{R}^{l_u}$ is the control input, $f : \mathbb{R}^n \to \mathbb{R}^n$ and $g : \mathbb{R}^n \to \mathbb{R}^n$ are nonlinear functions that represent the IEEE13NTF model, $d(k) \in \mathbb{R}^{l_d}$ is the disturbance, $y_0(k) \in \mathbb{R}^{l_y}$ is the plant output, $y_n(k) \in \mathbb{R}^{l_y}$ is the measurement, $r(k) \in \mathbb{R}^{l_y}$ is the command, $v(k) \in \mathbb{R}^{l_y}$ is the sensor noise, and $z(k) \in \mathbb{R}^{l_y}$ is the measured error, which is also the performance variable. Note that l_u, l_d, and l_y are the dimensions that represent the control, disturbance, and measurement vectors, respectively. The model in (1)–(4) can be viewed as a synchronously sampled, sampled-data version of IEEE13NTF shown in Figure 1, where the sensing, actuation, and load/generation disturbances are on physically separated nodes. IEEE13NTF is a quasi-static model that requires the iterative solution of a set of nonlinear algebraic equations. Therefore, the algebraic

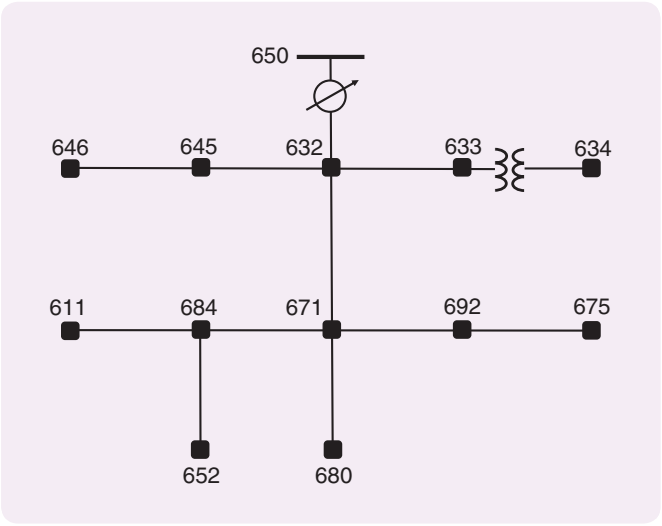


FIGURE 1 The IEEE 13-node test-feeder model. The model used for this article is provided by the OPAL-RT RT-Lab ePHASORSIM package.

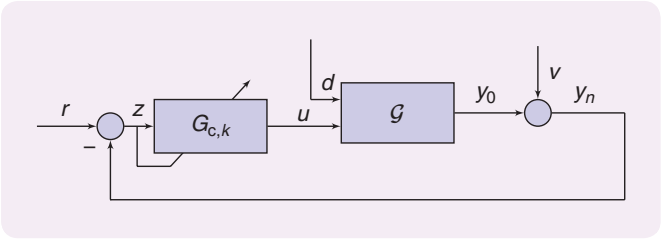


FIGURE 2 A block diagram representation of the adaptive servo problem with the adaptive controller $G_{c,k}$ and IEEE 13-node test feeder $\mathcal{G}$.

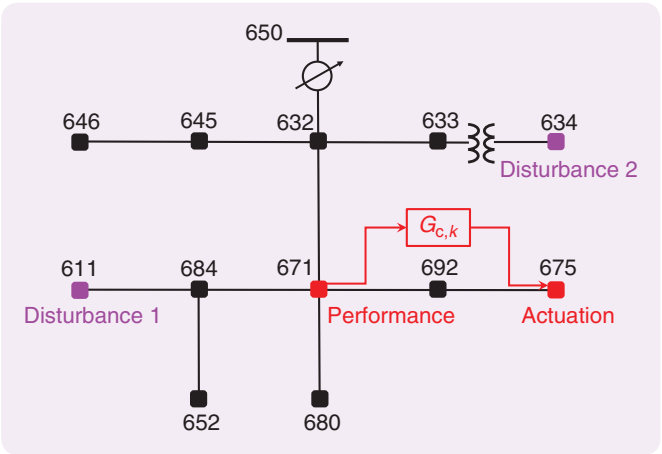


FIGURE 3 Nominal simulation tuning: at node 675, the controller $G_{c,k}$ can add or remove active and reactive power. At node 671, the controller $G_{c,k}$ has access to the voltage magnitude and angle measurements, where the voltage angles are relative to the slack bus at node 650. At nodes 611 and 634, the active and reactive power is varied, which represents an unmodeled load disturbance.

Fixed-Structure Retrospective Cost Adaptive Control Algorithm

CONTROLLER STRUCTURE

Define the dynamic compensator

$$u(k) = \Phi(k)\theta(k), \tag{S1}$$

where $\Phi(k) \in \mathbb{R}^{l_u \times l_\theta}$ is the regressor matrix constructed from past $y(k)$ and possibly past $u(k)$ data, and $\theta(k) \in \mathbb{R}^{l_\theta}$ is the controller coefficient vector.

RETROSPECTIVE PERFORMANCE

Define the retrospective control

$$\hat{u} \triangleq \Phi(k)\hat{\theta}, \tag{S2}$$

where $\hat{\theta} \in \mathbb{R}^{l_\theta}$ is determined by minimizing the retrospective cost function (S8). Using the retrospective control, define the retrospective performance variable as

$$\hat{z}(k,\hat{\theta}) \triangleq z(k) - G_{\mathrm{f}}(\mathbf{q})[u(k) - \hat{u}(k)], \tag{S3}$$

where $z(k), \hat{z}(k) \in \mathbb{R}^{l_z}$ and $u(k), \hat{u}(k) \in \mathbb{R}^{l_u}$. The optimal controller coefficient vector $\hat{\theta}_{\mathrm{opt}}(k)$, which is obtained by retrospective optimization (explained in the following section), yields the updated controller with coefficients $\theta(k+1) = \hat{\theta}_{\mathrm{opt}}(k)$ to be used at the next time step.

The $n_z \times n_u$ finite-impulse-response filter G_{f} of order n_{f} has the form

$$G_{\mathrm{f}}(\mathbf{q}) \triangleq \sum_{i=1}^{n_{\mathrm{f}}} N_i \mathbf{q}^{-i}, \tag{S4}$$

where N_i is an $l_z \times l_u$ matrix and $\mathbf{q}$ is the forward-shift operator. Equation (S3) can be rewritten as

$$\hat{z}(k,\hat{\theta}) \triangleq z(k) - N[\bar{U}(k) - \bar{\Phi}(k)\hat{\theta}], \tag{S5}$$

where

$$N \triangleq [N_1 \ \cdots \ N_{n_{\mathrm{f}}}] \tag{S6}$$

$$\bar{\Phi}(k) \triangleq \begin{bmatrix} \Phi(k-1) \\ \vdots \\ \Phi(k-n_{\mathrm{f}}) \end{bmatrix}, \quad \bar{U}(k) \triangleq \begin{bmatrix} u(k-1) \\ \vdots \\ u(k-n_{\mathrm{f}}) \end{bmatrix}, \tag{S7}$$

where $N \in \mathbb{R}^{l_z \times n_{\mathrm{f}} l_u}$, $\bar{\Phi}(k) \in \mathbb{R}^{n_{\mathrm{f}} l_u \times l_\theta}$, and $\bar{U}(k) \in \mathbb{R}^{n_{\mathrm{f}} l_u}$.

RETROSPECTIVE COST

Using the retrospective performance variable $\hat{z}(k,\hat{\theta})$ defined by (S3), define the cumulative retrospective cost function as

$$J(k,\hat{\theta}) \triangleq \sum_{i=1}^{k} \hat{z}^{\mathrm{T}}(i,\hat{\theta}) R_z \hat{z}(i,\hat{\theta}) + \sum_{i=1}^{k} (\Phi(i)\hat{\theta})^{\mathrm{T}} \bar{R}_u \Phi(i)\hat{\theta} + \hat{\theta}^{\mathrm{T}} \bar{R}_\theta \hat{\theta}, \tag{S8}$$

where $\bar{R}_\theta$ is positive definite, R_z is positive definite, and $\bar{R}_u$ is positive semidefinite. Recursive minimization of (S8) is used to update the controller coefficient vector $\hat{\theta}$. The following result uses recursive least squares to obtain the minimizer of (S8).

Proposition

Let $P(0) = \bar{R}_\theta^{-1}$. Then, for all $k \geq 1$, the retrospective cost function (S8) has the unique global minimizer $\hat{\theta}_{\mathrm{opt}}(k)$, which is given by

$$P(k) = P(k-1)[I_{l_\theta} - \gamma^{\mathrm{T}}(k)\beta(k)\gamma(k)P(k-1)], \tag{S9}$$

$$\hat{\theta}_{\mathrm{opt}}(k) = \hat{\theta}_{\mathrm{opt}}(k-1) - P(k)\gamma^{\mathrm{T}}(k)\bar{R}\psi(k), \tag{S10}$$

where

$$\beta(k) \triangleq [\bar{R}^{-1} + \gamma(k)P(k-1)\gamma^{\mathrm{T}}(k)]^{-1}, \tag{S11}$$

$$\gamma(k) \triangleq \begin{bmatrix} N\bar{\Phi}(k) \\ \Phi(k) \end{bmatrix}, \quad \bar{R} \triangleq \begin{bmatrix} R_z & 0_{l_z \times l_u} \\ 0_{l_u \times l_z} & \bar{R}_u \end{bmatrix}, \tag{S12}$$

$$\psi(k) \triangleq \begin{bmatrix} N\bar{\Phi}(k)\hat{\theta}_{\mathrm{opt}}(k-1) + z(k) - N\bar{U}(k) \\ \Phi(k)\hat{\theta}_{\mathrm{opt}}(k-1) \end{bmatrix}. \tag{S13}$$

The implemented control at step $k+1$ is given by

solution of (1)–(4) determines the state update and effective impulse response.

ADAPTIVE CONTROL ALGORITHM

Despite the fact that the model in (1)–(4) is a nonlinear, time-varying plant, RCAC uses extremely limited information to update the controller coefficients. Furthermore, no explicit information about the nonlinearity or time variation is used to select the adaptive tuning weights. The goal of this study was to determine the extent to which RCAC can accommodate the unmodeled features of (1)–(4). For SISO linear plants, RCAC requires limited modeling data, namely, the sign of the leading numerator coefficient, the relative degree, and the locations of NMP zeros, if any. This information is used to construct the filter G_{f}, which is used to compute the retrospective performance variable. As explained in [32], G_{f} serves as a target model for the closed-loop intercalated transfer function from the virtual control input perturbation to the performance variable. For application to IEEE13NTF, the RCAC tuning weights are chosen based on a combination of NST and PST.

At each step k, RCAC updates the feedback controller $G_{\mathrm{c},k}$ to optimally fit the intercalated transfer function to the target model G_{f}. The input of $G_{\mathrm{c},k}$ is the three voltage magnitude errors and three voltage angle errors at the performance node. Its output is the three active powers and three reactive powers extracted at the actuation node. The controller order and structure are chosen by the user, as are the adaptation weight R_θ, control weight R_u, and

$$u(k+1) = \Phi(k+1)\hat{\theta}_{\text{opt}}(k) = \Phi(k+1)\theta(k+1), \tag{S14}$$

where

$$\theta(k+1) \triangleq \hat{\theta}_{\text{opt}}(k). \tag{S15}$$

$\hat{\theta}_{\text{opt}}(0) = 0$ is initialized to reflect the absence of additional prior modeling information, and $\bar{R}_u = R_u I_{l_u}$ and $\bar{R}_\theta = R_\theta I_{l_\theta}$ are used where R_u and R_θ are scalar weightings.

FIXED-STRUCTURE CONTROLLER PARAMETERIZATION

Define the dynamic compensator of the form

$$u(k) = G_{\text{c},k}(\mathbf{q})y(k), \tag{S16}$$

$$G_{\text{c},k} \triangleq \begin{bmatrix} G_{\text{c},11,k} & \cdots & G_{\text{c},1l_y,k} \\ \vdots & \ddots & \vdots \\ G_{\text{c},l_u1,k} & \cdots & G_{\text{c},l_ul_y,k} \end{bmatrix}, \tag{S17}$$

where each entry $G_{\text{c},ij,k}$ of the $l_u \times l_y$ discrete-time transfer function $G_{\text{c},k}(\mathbf{q})$ is either an infinite-impulse-response (IIR) controller of a specified order ℓ_{ij} or a proportional-integral (PI) controller. That is,

$$G_{\text{c},ij,k}(\mathbf{q}) = \begin{cases} \dfrac{q_{ij,1}(k)\mathbf{q}^{\ell_{ij}-1} + \cdots + q_{ij,\ell_{ij}}(k)}{\mathbf{q}^{\ell_{ij}} - p_{ij,1}(k)\mathbf{q}^{\ell_{ij}-1} - \cdots - p_{ij,\ell_{ij}}(k)}, & \text{IIR}, \\ \dfrac{q_{ij,1}(k)}{\mathbf{q}} + \dfrac{q_{ij,2}(k)}{\mathbf{q}-1}, & \text{PI}. \end{cases} \tag{S18}$$

To construct $\theta(k)$, first define

$$\theta_{ij}(k) \triangleq \begin{cases} \begin{bmatrix} p_{ij,1}(k) \\ \vdots \\ p_{ij,\ell_{\text{c},ij}}(k) \\ q_{ij,1}(k) \\ \vdots \\ q_{ij,\ell_{\text{c},ij}}(k) \end{bmatrix} \in \mathbb{R}^{l_{\theta,ij}}, & \text{IIR } G_{\text{c},ij,k}, \\ \begin{bmatrix} q_{ij,1}(k) \\ q_{ij,2}(k) \end{bmatrix} \in \mathbb{R}^2, & \text{PI } G_{\text{c},ij,k}, \end{cases} \tag{S19}$$

which, in turn, are used to define

$$\theta_i(k) \triangleq \begin{bmatrix} \theta_{i1} \\ \vdots \\ \theta_{il_y} \end{bmatrix} \in \mathbb{R}^{l_{\theta_i}}, \tag{S20}$$

and, finally, define

$$\theta(k) \triangleq \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_{l_u} \end{bmatrix} \in \mathbb{R}^{l_\theta}, \tag{S21}$$

where

$$l_{\theta,ij} = \begin{cases} 2\ell_{\text{c},ij}, & \text{IIR } G_{\text{c},ij,k}, \\ 2, & \text{PI } G_{\text{c},ij,k}, \end{cases} \tag{S22}$$

$$l_{\theta_i} \triangleq \sum_{j=1}^{l_y} l_{\theta,ij}, \quad l_\theta \triangleq \sum_{i=1}^{l_u}\sum_{j=1}^{l_y} l_{\theta,ij}. \tag{S23}$$

For each controller entry $G_{\text{c},ij,k}$, define

$$u_{ij}(k) \triangleq G_{\text{c},ij,k}(\mathbf{q})y_j(k), \tag{S24}$$

and construct $\phi_{ij}(k)$. For an IIR $G_{\text{c},ij,k}$ of order ℓ_{ij},

$$\phi_{ij}(k) \triangleq \begin{bmatrix} u_{ij}(k-1) \\ \vdots \\ u_{ij}(k-\ell_{ij}) \\ y_j(k-1) \\ \vdots \\ y_j(k-\ell_{ij}) \end{bmatrix}^T \in \mathbb{R}^{1\times l_{\theta,ij}}, \tag{S25}$$

and for a PI $G_{\text{c},ij,k}$,

$$\phi_{ij}(k) \triangleq \begin{bmatrix} y_j(k-1) \\ \sum_{p=0}^{k-1} y_j(p) \end{bmatrix}^T \in \mathbb{R}^{1\times l_{\theta,ij}}. \tag{S26}$$

Using $\phi_{ij}(k)$, construct

$$\phi_i(k) \triangleq [\phi_{i1} \ \cdots \ \phi_{il_y}] \in \mathbb{R}^{1\times l_{\theta_i}}, \tag{S27}$$

which in turn is used to construct

$$\Phi(k) \triangleq \text{diag}(\phi_1(k), \ldots, \phi_{l_u}(k)) \in \mathbb{R}^{l_u \times l_\theta}. \tag{S28}$$

cost weight R_z. The coefficients of $G_{\text{c},k}$ are entries of the controller coefficient vector $\theta(k)$. As in [32], the initial value $\theta(0)$ of $\theta(k)$ is set to zero for all simulations. This assumption reflects the absence of additional modeling information. In practice, $\theta(0)$ can be chosen to be nonzero based on modeling or prior adaptation. Details of RCAC and its implementation using recursive least squares to update θ are given in [32]. A summary of the algorithm is presented in "Fixed-Structure Retrospective Cost Adaptive Control Algorithm."

PROBLEM SETUP

Consider IEEE13NTF from Figure 1 (which is simulated with OpalRT RT-LAB, version 2017.0.4.59) and use the fixed-structure RCAC described in "Fixed-Structure Retrospective Cost Adaptive Control Algorithm" with a sparse target model G_{f}. RCAC is updated at 100 Hz, which is sufficiently fast for the time-varying load and PV-generation dynamics considered in this study. In this article, real customer PV-generation data are obtained from Pecan Street [36] and integrated into IEEE13NTF. In what follows, continuous time t and the sampled time step k are related by $t = k/100$.

In this article, only three phase performance nodes for command-following are considered. Specifically, for each performance node, define

$$y_0(k) \triangleq \begin{bmatrix} y_{0,\text{mag}}(k) \\ y_{0,\text{ang}}(k) \end{bmatrix}, \quad z(k) \triangleq \begin{bmatrix} z_{\text{mag}}(k) \\ z_{\text{ang}}(k) \end{bmatrix}, \tag{5}$$

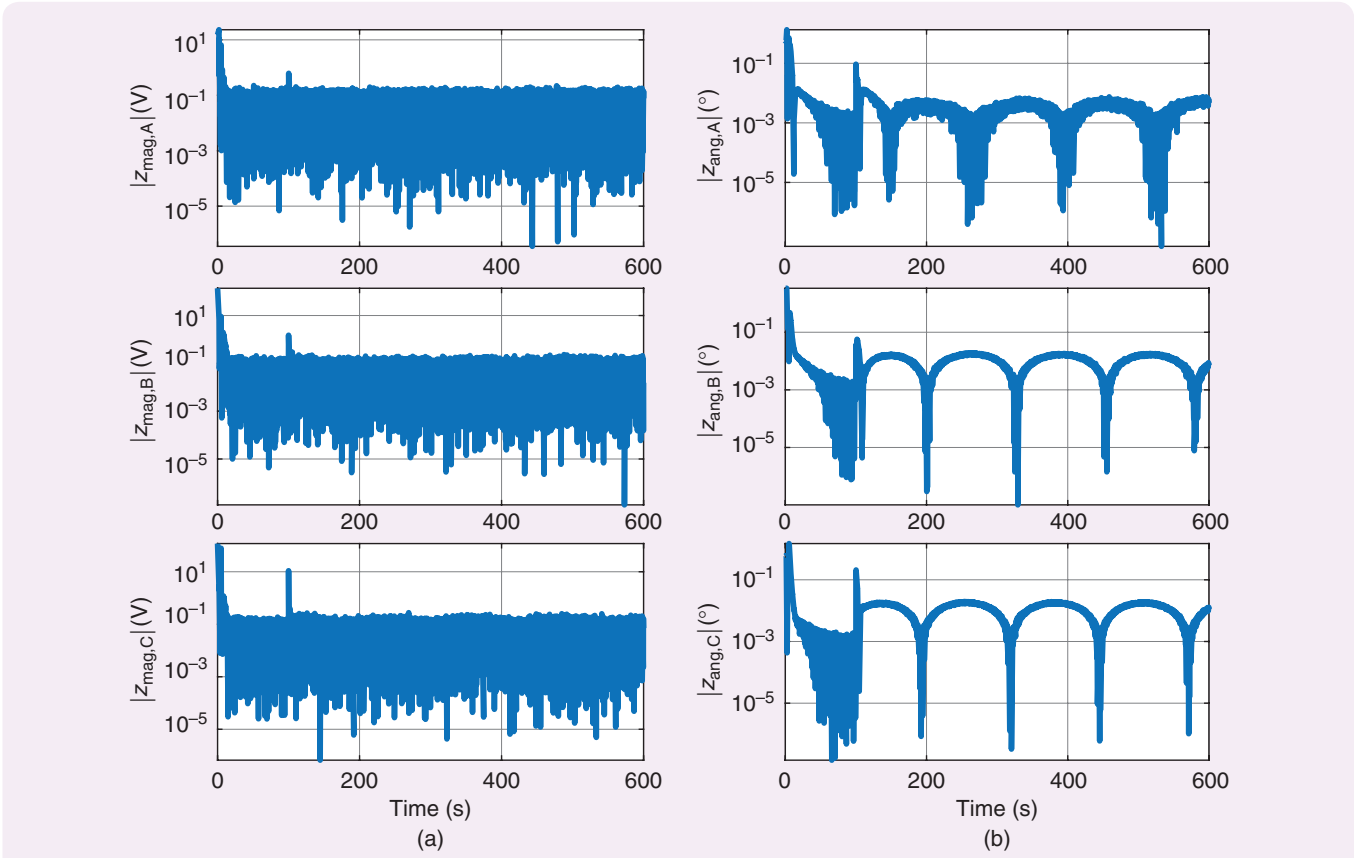


FIGURE 4 Nominal simulation tuning: the (a) voltage-magnitude error $|z_{mag}|$ and (b) voltage angle-magnitude error $|z_{ang}|$ at node 671 are shown on a logarithmic scale. Asymptotically, the voltage-magnitude and voltage-angle errors are approximately lower than 0.1 V and 0.01°, respectively.

where

$$y_{0,\mathrm{mag}}(k) \triangleq \begin{bmatrix} V_{\mathrm{mag,A}}(k) \\ V_{\mathrm{mag,B}}(k) \\ V_{\mathrm{mag,C}}(k) \end{bmatrix}, \quad y_{0,\mathrm{ang}}(k) \triangleq \begin{bmatrix} V_{\mathrm{ang,A}}(k) \\ V_{\mathrm{ang,B}}(k) \\ V_{\mathrm{ang,C}}(k) \end{bmatrix}, \tag{6}$$

$$z_{\mathrm{mag}}(k) \triangleq \begin{bmatrix} z_{\mathrm{mag,A}}(k) \\ z_{\mathrm{mag,B}}(k) \\ z_{\mathrm{mag,C}}(k) \end{bmatrix}, \quad z_{\mathrm{ang}}(k) \triangleq \begin{bmatrix} z_{\mathrm{ang,A}}(k) \\ z_{\mathrm{ang,B}}(k) \\ z_{\mathrm{ang,C}}(k) \end{bmatrix}. \tag{7}$$

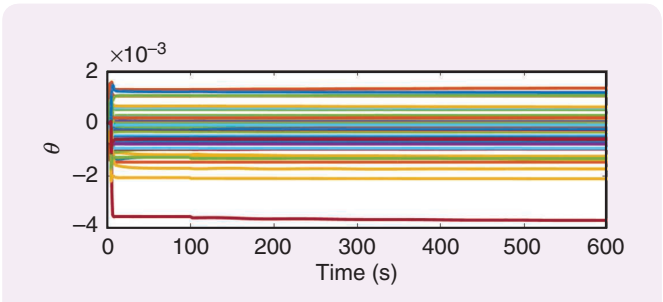


FIGURE 5 Nominal simulation tuning: the retrospective cost adaptive control (RCAC) controller $G_{c,k}$ adapts to follow the setpoint commands. RCAC readapts at $t = 5$ s to account for the step-load disturbance and again at $t = 100$ s to account for the additional harmonic component of the load disturbance.

Phasor-based control is set up as a servo problem, as shown in Figure 2, with the objective being phase balancing. Consequently, r has the same dimensions as y_0 and is defined as

$$r(k) \triangleq \begin{bmatrix} r_{\mathrm{mag}} \\ r_{\mathrm{mag}} \\ r_{\mathrm{mag}} \\ r_{\mathrm{ang}} \\ r_{\mathrm{ang}} - 120^\circ \\ r_{\mathrm{ang}} + 120^\circ \end{bmatrix}, \tag{8}$$

which allows for specification of the command for each node as a pair of two scalars $r_{\mathrm{mag}}, r_{\mathrm{ang}}$ instead of six scalars. This reference is designed to balance the three phases.

Actuators are modeled as idealized three-phase, four-quadrant operation dc/ac inverters, which are connected to infinitely large battery packs at the actuation nodes. To avoid the case in which an actuation node goes offline, we assume that each battery pack has infinite capacity. Each RCAC controller commands a single inverter, which can add or extract active and reactive power at each of the three phases. The output $u(k)$ of the controller $G_{c,k}$ at step k is defined to be $u(k)$, scaled by a fixed scaling. An effective scaling is determined from the simulated closed-loop response. Specifically, the power extracted at an actuation node is given by

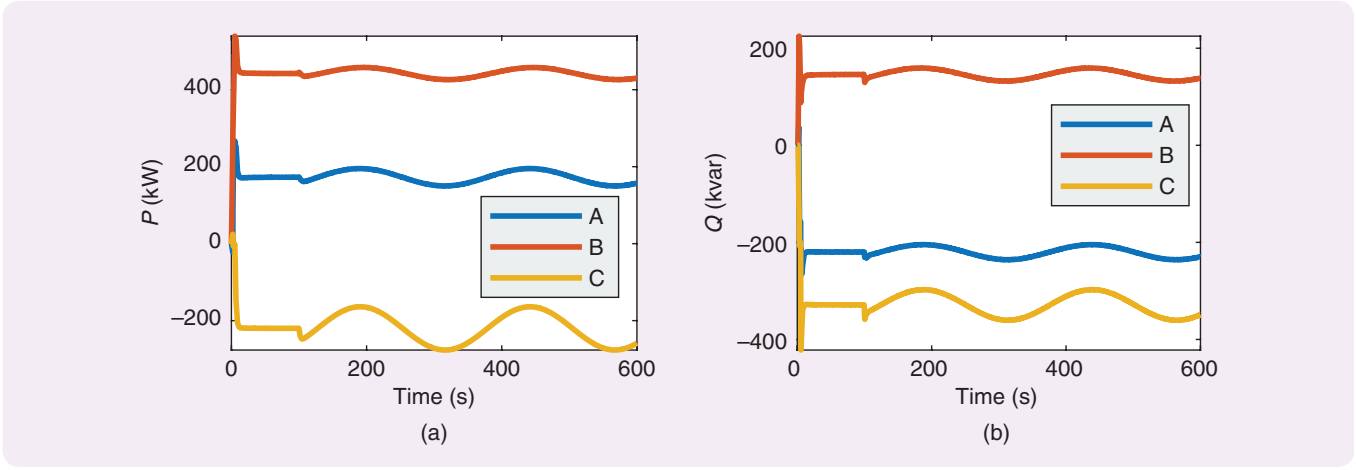


FIGURE 6 Nominal simulation tuning: the (a) active and (b) reactive power extracted from each phase by the retrospective cost adaptive control controller $G_{c,k}$ at node 675 is shown.

$$P_{\text{act}}(k) \triangleq \begin{bmatrix} P_{\text{act},A}(k) \\ Q_{\text{act},A}(k) \\ P_{\text{act},B}(k) \\ Q_{\text{act},B}(k) \\ P_{\text{act},C}(k) \\ Q_{\text{act},C}(k) \end{bmatrix} = 200u(k). \tag{9}$$

Since only three-phase performance and actuation nodes are considered, it follows that $l_u = l_y = 6$ for all simulations. Additionally, define

$$v(k) \triangleq \begin{bmatrix} v_{\text{mag}}(k) \\ v_{\text{ang}}(k) \end{bmatrix}, \tag{10}$$

$$v_{\text{mag}}(k) \triangleq \begin{bmatrix} v_{\text{mag},A}(k) \\ v_{\text{mag},B}(k) \\ v_{\text{mag},C}(k) \end{bmatrix}, \quad v_{\text{ang}}(k) \triangleq \begin{bmatrix} v_{\text{ang},A}(k) \\ v_{\text{ang},B}(k) \\ v_{\text{ang},C}(k) \end{bmatrix}, \tag{11}$$

where each component of $v_{\text{mag}}(k)$ is Gaussian white noise with standard deviation 0.05 V, and each component of $v_{\text{ang}}(k)$ is Gaussian white noise with standard deviation 0.0005°.

NOMINAL SIMULATION TUNING

In this section, the controller structure, weightings, and target model are determined by running multiple simulations of a single scenario. The scenario selected for NST is command-following at node 671, with load disturbances at nodes 611 and 634 and actuation at node 675 (Figure 3).

Define the active and reactive power extracted at node 634 as

$$\begin{bmatrix} P_{634,A}(k) \\ Q_{634,A}(k) \\ P_{634,B}(k) \\ Q_{634,B}(k) \\ P_{634,C}(k) \\ Q_{634,C}(k) \end{bmatrix} = (1 + d(k)) \begin{bmatrix} 160\ \text{kW} \\ 110\ \text{kvar} \\ 120\ \text{kW} \\ 90\ \text{kvar} \\ 120\ \text{kW} \\ 90\ \text{kvar} \end{bmatrix}, \tag{12}$$

the active and reactive power extracted at node 611 as

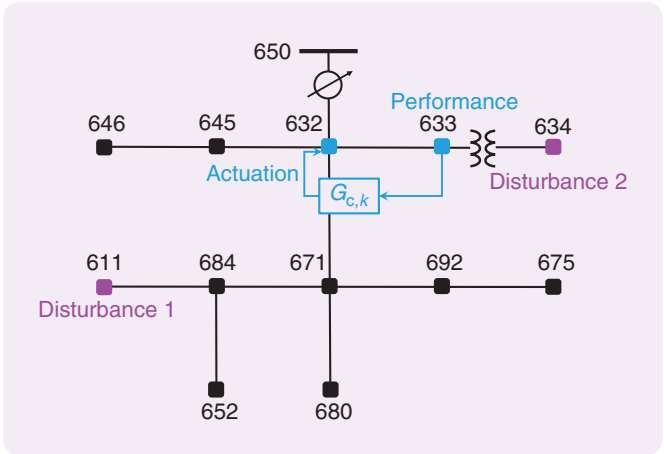


FIGURE 7 Perturbed simulation testing 1: at node 632, the controller $G_{c,k}$ can add or remove active and reactive power. At node 633, the controller $G_{c,k}$ has access to the voltage magnitude and angle measurements, where the voltage angles are relative to the slack bus at node 650. At nodes 611 and 634, the active and reactive power is varied, which represents an unmodeled load disturbance.

$$\begin{bmatrix} P_{611}(k) \\ Q_{611}(k) \end{bmatrix} = (1 + d(k)) \begin{bmatrix} 170\ \text{kW} \\ 80\ \text{kvar} \end{bmatrix}, \tag{13}$$

and the time-varying load disturbance $d(k)$ as

$$d(k) = \begin{cases} 0, & k < 500, \\ 1, & 500 \le k < 10000, \\ 1 + \dfrac{\sin 0.00025k}{4}, & k \ge 10000, \end{cases} \tag{14}$$

and set

$$r_{\text{mag}}(k) = 2267\ V \tag{15}$$

$$r_{\text{ang}}(k) = -5.3°. \tag{16}$$

To simplify the MIMO controller, each SISO entry of the MIMO controller is chosen to be either PI or second-order IIR. Based on simulation (5), the 6×6 controller structure

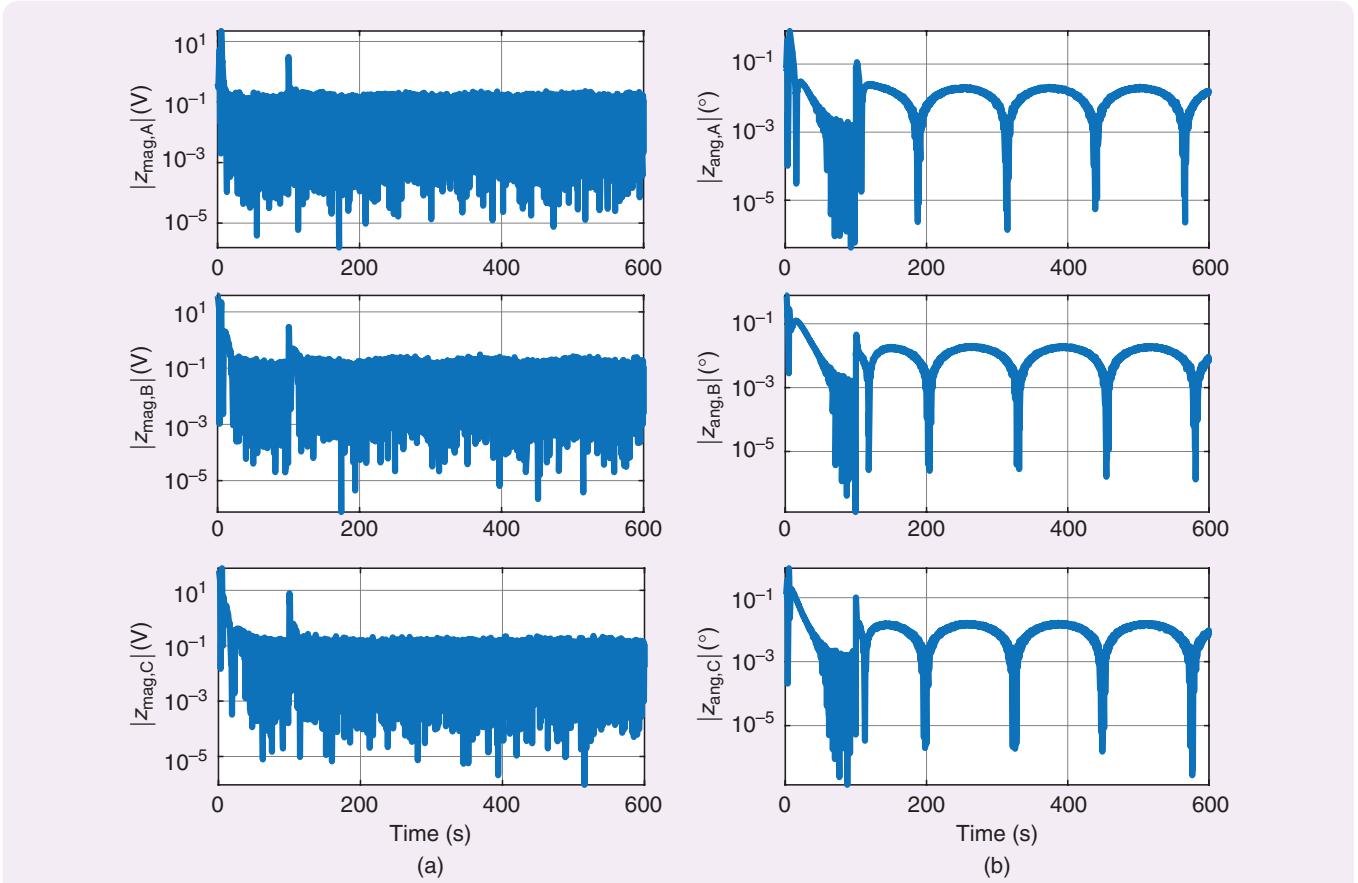


FIGURE 8 Perturbed simulation testing 1: the (a) voltage-magnitude error $|z_{mag}|$ and (b) voltage angle-magnitude error $|z_{ang}|$ at node 633 are shown on a logarithmic scale. Asymptotically, the voltage-magnitude and voltage-angle errors are approximately lower than 0.1 V and 0.01°, respectively.

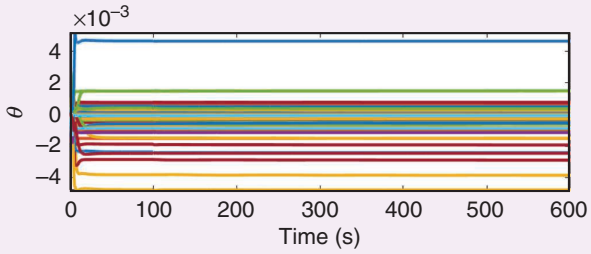


FIGURE 9 Perturbed simulation testing 1: the retrospective cost adaptive control (RCAC) controller $G_{c,k}$ adapts to follow the set-point commands. RCAC readapts at $t = 5$ s to account for the step-load disturbance and again at $t = 100$ s to account for the additional harmonic component of the load disturbance.

with the least complexity that follows phase-balancing commands is given by

$$\begin{bmatrix} \text{PI} & \text{IIR} & \text{PI} & \text{PI} & \text{PI} & \text{PI} \\ \text{PI} & \text{PI} & \text{PI} & \text{IIR} & \text{PI} & \text{PI} \\ \text{PI} & \text{PI} & \text{PI} & \text{PI} & \text{PI} & \text{IIR} \\ \text{IIR} & \text{PI} & \text{PI} & \text{PI} & \text{PI} & \text{PI} \\ \text{PI} & \text{PI} & \text{IIR} & \text{PI} & \text{PI} & \text{PI} \\ \text{PI} & \text{PI} & \text{PI} & \text{PI} & \text{IIR} & \text{PI} \end{bmatrix}. \tag{17}$$

Fixed-structure RCAC, described in "Fixed-Structure Retrospective Cost Adaptive Control Algorithm," is used to adapt the controller coefficients. For (17), RCAC adapts 84 coefficients. If each entry of (17) were chosen to be second-order IIR, then RCAC would need to adapt 144 coefficients.

Additionally, set $R_\theta = 10^{-9}$, $R_u = 0$, and $R_z = \text{diag}(1, 1, 1, 100, 100, 100)$ and use $G_f(\mathbf{q}) = (1/\mathbf{q})\hat{H}_1$, where $\hat{H}_1$ is constructed by the following procedure:

1) First, obtain H_1 by impulsing each input at the actuation node in IEEE13NTF at the initial conditions and recording the first value of each measurement at the performance node. Separately replace each entry of H_1 with zero, and record the entries whose replacement by zero does not cause RCAC to fail. These entries are highlighted in bold:

$$\begin{bmatrix} \mathbf{92.54} & 67.93 & \mathbf{6.71} & \mathbf{-8.39} & \mathbf{-48.21} & \mathbf{-3.84} \\ \mathbf{-89.25} & \mathbf{-3.81} & \mathbf{9.81} & 20.64 & \mathbf{14.38} & \mathbf{-28.56} \\ \mathbf{24.75} & \mathbf{-26.95} & \mathbf{-9.35} & \mathbf{-0.12} & \mathbf{58.68} & 83.23 \\ 4.03 & \mathbf{-0.50} & \mathbf{-0.18} & \mathbf{-0.15} & \mathbf{0.07} & \mathbf{0.87} \\ \mathbf{0.12} & \mathbf{0.77} & 0.48 & \mathbf{-0.20} & \mathbf{-0.97} & \mathbf{-0.47} \\ -1.74 & -0.52 & \mathbf{-0.07} & \mathbf{0.15} & 2.60 & -0.66 \end{bmatrix}.$$

2) Next, separately replace each entry of H_1 with its additive inverse, and record the entries for which

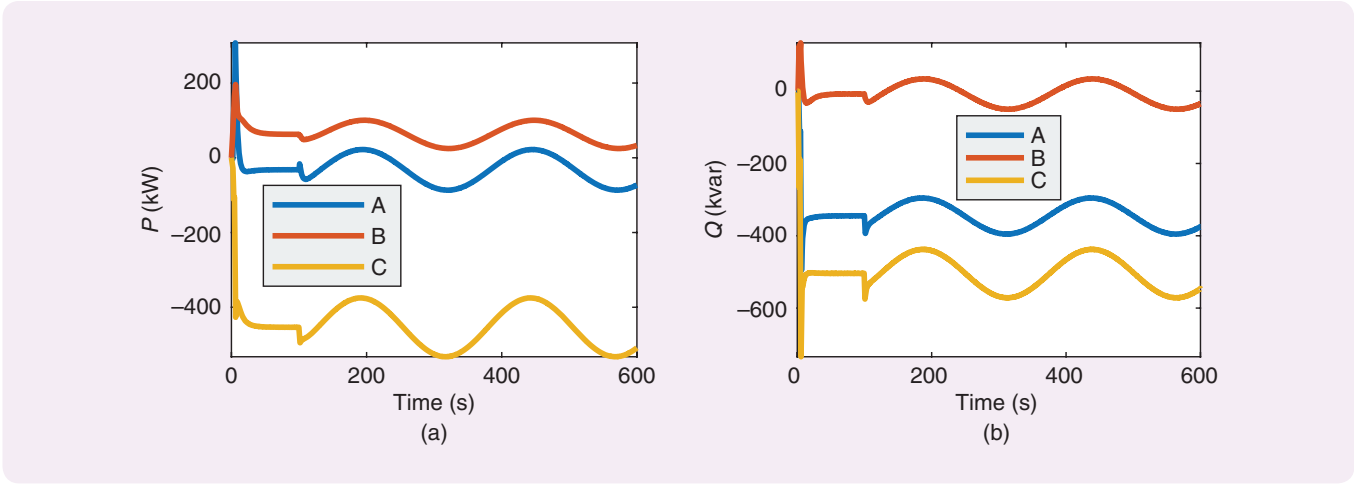


FIGURE 10 Perturbed simulation testing 1: the (a) active and (b) reactive power extracted from each phase by the retrospective cost adaptive control controller $G_{c,k}$ at node 632 is shown.

replacement by the additive inverse does not cause RCAC to fail. These entries are highlighted in bold:

$$\begin{bmatrix} \mathbf{92.54} & 67.93 & 6.71 & \mathbf{-8.39} & \mathbf{-48.21} & \mathbf{-3.84} \\ \mathbf{-89.25} & \mathbf{-3.81} & \mathbf{9.81} & 20.64 & \mathbf{14.38} & \mathbf{-28.56} \\ \mathbf{24.75} & -26.95 & -9.35 & \mathbf{-0.12} & \mathbf{58.68} & 83.23 \\ 4.03 & \mathbf{-0.50} & \mathbf{-0.18} & \mathbf{-0.15} & \mathbf{0.07} & \mathbf{0.87} \\ \mathbf{0.12} & \mathbf{0.77} & 0.48 & \mathbf{-0.20} & \mathbf{-0.97} & \mathbf{-0.47} \\ \mathbf{-1.74} & \mathbf{-0.52} & \mathbf{-0.07} & \mathbf{0.15} & 2.60 & -0.66 \end{bmatrix}.$$

3) Replace all 29 of the bold entries in steps 1 and 2 with zero, which yields

$$\begin{bmatrix} 0 & 67.93 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 20.64 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 83.23 \\ 4.03 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.48 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2.60 & -0.66 \end{bmatrix}.$$

4) Next, separately replace each nonzero entry in $\hat{H}_1$ constructed in the previous step with zero to determine the modified matrix $\hat{H}_1$ that has the smallest number of nonzero entries for which RCAC does not fail, which is given by

$$\begin{bmatrix} 0 & 67.93 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 20.64 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 83.23 \\ 4.03 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.48 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2.60 & 0 \end{bmatrix}. \tag{18}$$

5) Finally, the numerical values in (18) suggest that it may be possible to parameterize (18) by using only two numbers. Numerical testing suggests 200 V for magnitude and 5° for angle, which yields

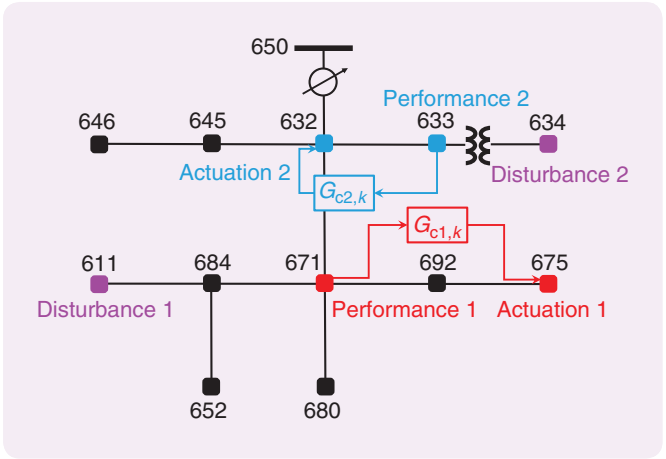


FIGURE 11 Perturbed simulation testing 2: at node 675, subcontroller $G_{c1,k}$ can add or remove active and reactive power. At node 671, subcontroller $G_{c1,k}$ has access to the voltage magnitude and angle measurements, where the voltage angles are relative to the slack bus at node 650. At node 632, subcontroller $G_{c2,k}$ can add or remove active and reactive power. At node 633, subcontroller $G_{c2,k}$ has access to the voltage magnitude and angle measurements, where the voltage angles are relative to the slack bus at node 650. At nodes 611 and 634, the active and reactive power is varied to represent an unmodeled load disturbance.

$$\hat{H}_1 = \begin{bmatrix} 0 & 200 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 200 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 200 \\ 5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 5 & 0 \end{bmatrix}. \tag{19}$$

Note that the rows of (19) can be rearranged to construct a 6×6 diagonal matrix. This rearrangement yields a specific pairing of the input and output variables in (5)–(7). Figures 4–6 illustrate RCAC using (19), the controller structure (17), and the chosen weightings, with voltage magnitude and angle commands at node 671, actuation at

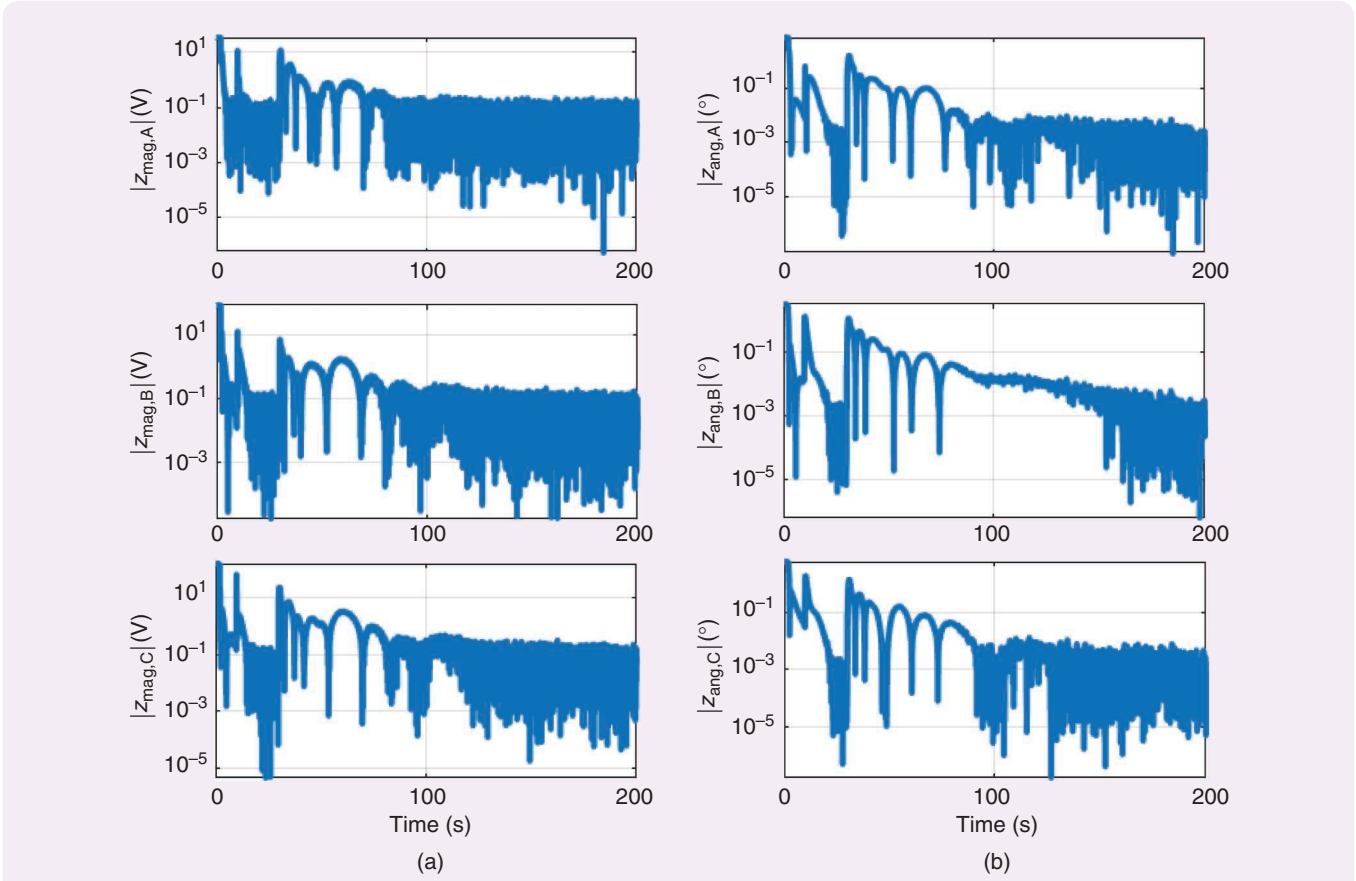


FIGURE 12 Perturbed simulation testing 2: the (a) voltage-magnitude error $|z_{mag}|$ and (b) voltage angle-magnitude error $|z_{ang}|$ at node 671 are shown on a logarithmic scale. Asymptotically, the voltage-magnitude and voltage-angle errors are approximately lower than 0.1 V and 0.01°, respectively.

node 675, and the time-varying load disturbance (14) at nodes 611 and 634 in the presence of the sensor noise (34). This completes the NST.

Perturbed Simulation Testing

In this section, the controller structure, weightings, and target model obtained by NST in the previous section are applied to multiple scenarios involving perturbed simulations. Each simulation represents a perturbation of the nominal simulation in terms of the disturbance signals; actuator saturation levels; and choice of disturbance, actuation, and performance nodes. For all simulations, the controller structure given by (17) is used with the weightings $R_\theta = 10^{-9}$, $R_u = 0$, and $R_z = \text{diag}(1, 1, 1, 100, 100, 100)$ and target model $G_f(\mathbf{q}) = (1/\mathbf{q})\hat{H}_1$, where $\hat{H}_1$ is given by (19). Each case is simulated once, and no attempt is made to retune RCAC based on the response of the closed-loop system.

Perturbed Simulation Testing 1 (PST 1)

PST 1 tests command-following at node 633, with load disturbance at nodes 611 and 634 and actuation at node 632 (Figure 7). The active and reactive power at nodes 634 and 611 is defined using (12) and (13), respectively, and the load disturbance $d(k)$ is set using (14). For node 633, the command is set as

$$r_{mag}(k) = 2312 \text{ V}, \tag{20}$$
$$r_{ang}(k) = -2.56°. \tag{21}$$

For the controller $G_{c,k}$, the tuning weights described at the beginning of this section are used. Feedback control with adaptation begins at $t = 2$ s. Figures 8–10 illustrate RCAC with voltage magnitude and angle commands at node 633, actuation at node 632, and the time-varying load disturbance (14) at nodes 611 and 634 in the presence of the sensor noise (34). ◊

Perturbed Simulation Testing 2 (PST 2)

PST 2 tests command-following at nodes 671 and 633, with load disturbances at nodes 611 and 634 and actuation at nodes 675 and 632, using decentralized control (Figure 11). The active and reactive power at nodes 634 and 611 is defined using (12) and (13), respectively, and the load disturbance $d(k)$ is set as

$$d(k) = \begin{cases} 0, & k < 1{,}000, \\ 1, & \text{otherwise.} \end{cases} \tag{22}$$

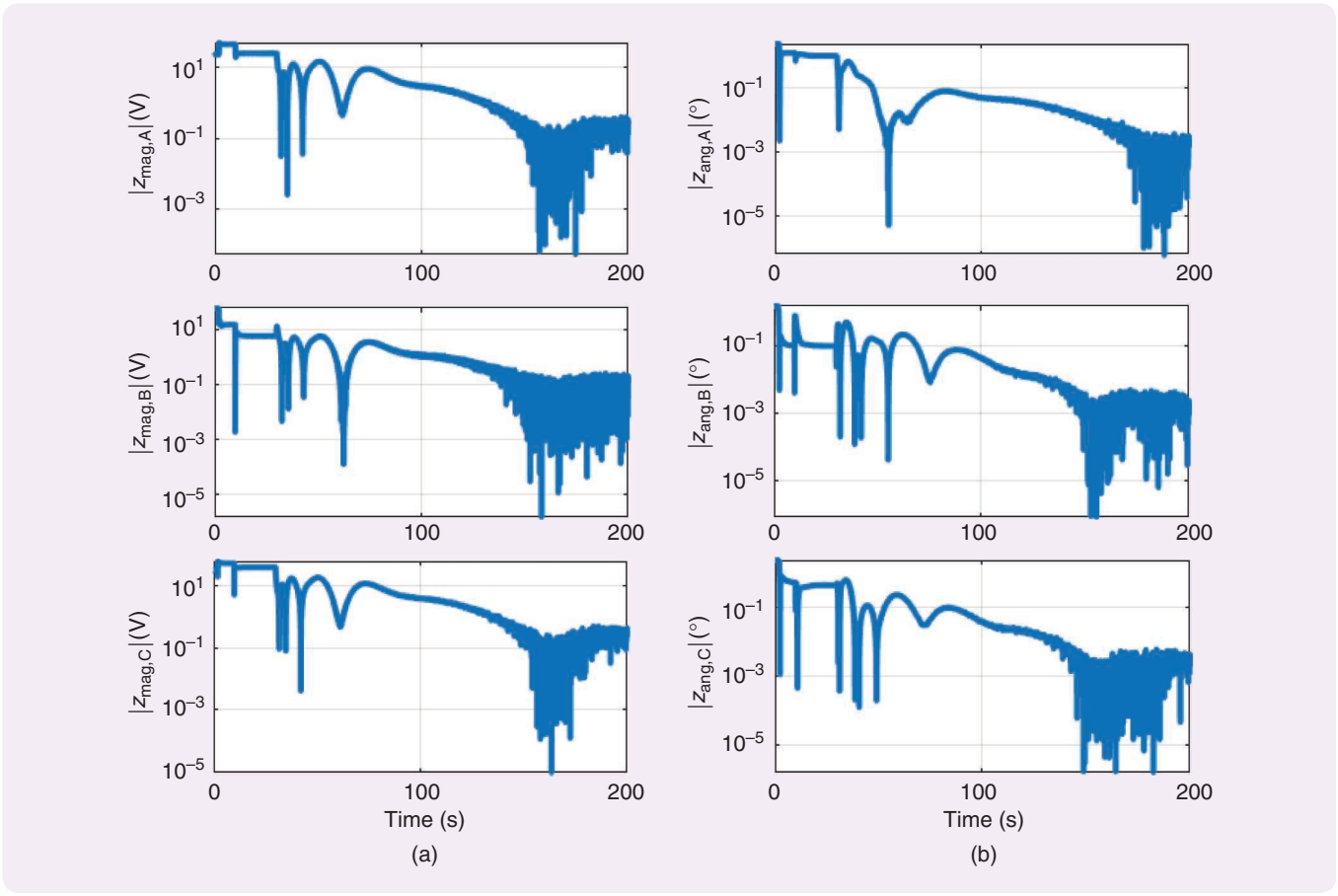


FIGURE 13 Perturbed simulation testing 2: the (a) voltage-magnitude error $|z_{mag}|$ and (b) voltage angle-magnitude error $|z_{ang}|$ for node 633 are shown on a logarithmic scale. Asymptotically, the voltage-magnitude and voltage-angle errors are approximately lower than 0.1 V and 0.01°, respectively.

For node 671, the command is set as

$$r_{671,\mathrm{mag}}(k) = 2{,}300\ \mathrm{V}, \tag{23}$$

$$r_{671,\mathrm{ang}}(k) = 1.23°, \tag{24}$$

and for node 633, the command is set as

$$r_{633,\mathrm{mag}}(k) = 2290\ \mathrm{V}, \tag{25}$$

$$r_{633,\mathrm{ang}}(k) = -0.34°. \tag{26}$$

For each subcontroller $G_{c1,k}$ and $G_{c2,k}$, the tuning weights described at the beginning of this section are used. Feedback control with adaptation using subcontrollers $G_{c1,k}$ and $G_{c2,k}$ begins at $t = 2$ s and $t = 30$ s, respectively. Figures 12–16 illustrate a pair of decentralized RCAC subcontrollers with voltage magnitude and angle commands at nodes 671 and 633, actuation at nodes 675 and 632, and the time-varying load disturbance (22) at nodes 611 and 634 in the presence of the sensor noise (34). There is no direct communication between subcontrollers $G_{c1,k}$ and $G_{c2,k}$. ◊

Perturbed Simulation Testing 3 (PST 3)

PST 3 tests command-following at node 671, with load disturbances at nodes 611 and 634 and actuation at nodes 675

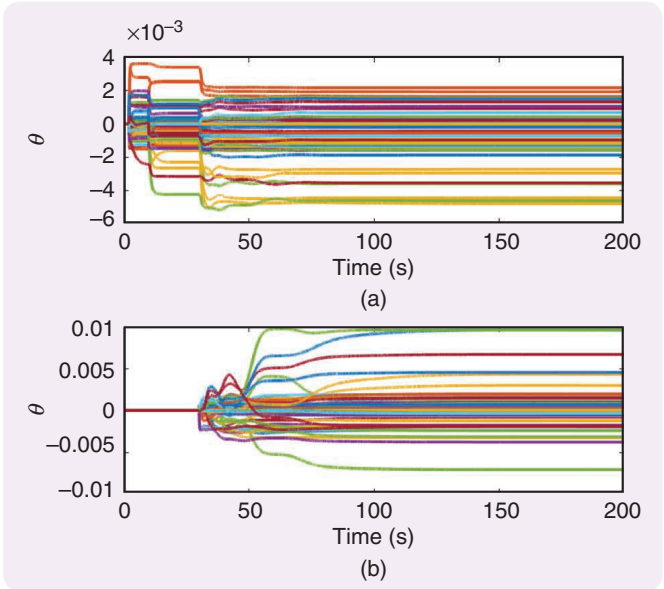


FIGURE 14 Perturbed simulation testing 2: (a) $\theta(k)$ for subcontroller $G_{c1,k}$ is zero for $t < 2$ s, after which the adaptation of subcontroller $G_{c1,k}$ begins to follow the setpoint commands. Subcontroller $G_{c1,k}$ readapts at $t = 10$ s to account for the step disturbance and again at $t = 30$ s to account for subcontroller $G_{c2,k}$ starting adaptation. (b) $\theta(k)$ for subcontroller $G_{c2,k}$ is zero for $t < 30$ s, after which adaptation of subcontroller $G_{c2,k}$ begins to follow the setpoint commands.

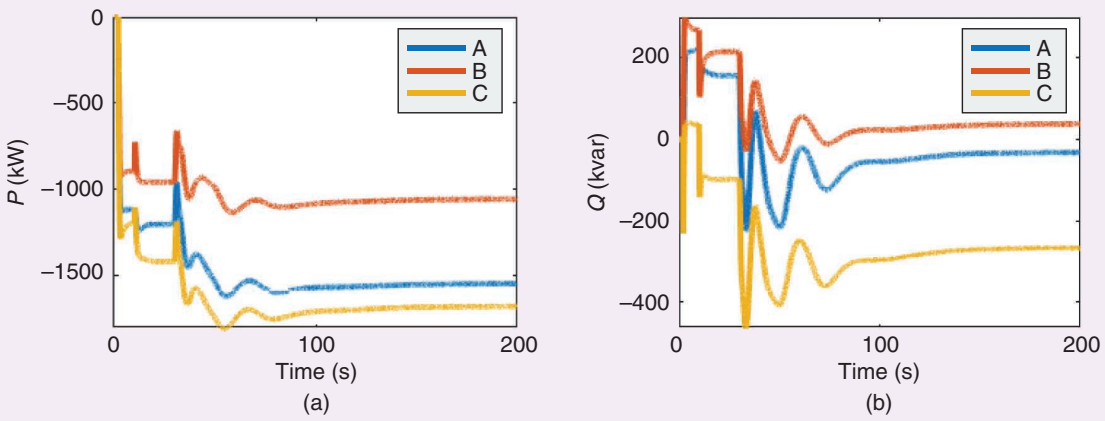


FIGURE 15 Perturbed simulation testing 2: the (a) active and (b) reactive power extracted by subcontroller $G_{c1,k}$ from each phase at node 675 is shown.

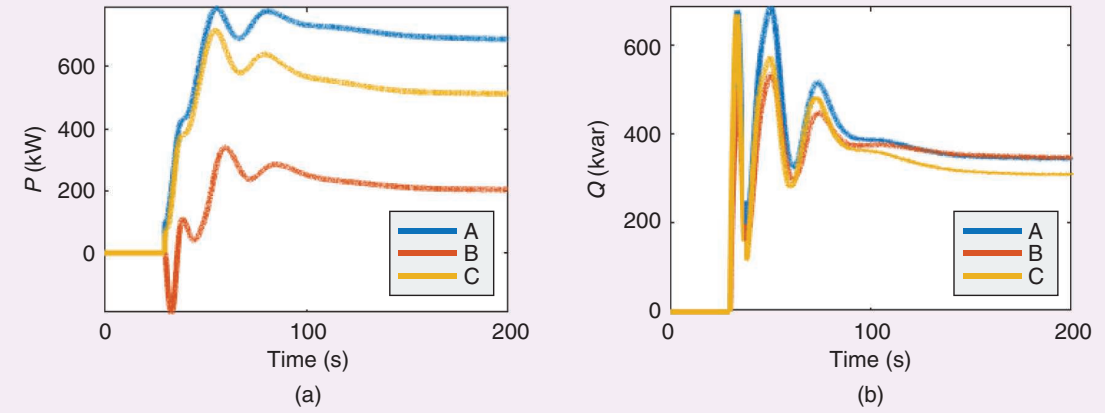


FIGURE 16 Perturbed simulation testing 2: the (a) active and (b) reactive power extracted by subcontroller $G_{c2,k}$ from each phase at node 632 is shown.

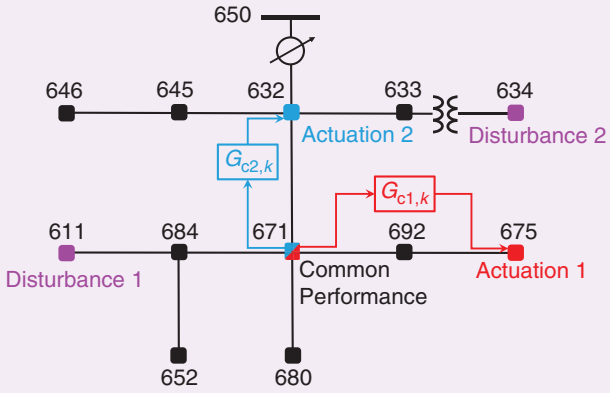


FIGURE 17 Perturbed simulation testing 3: at node 675 and node 632, subcontroller $G_{c1,k}$ and subcontroller $G_{c2,k}$ can add or remove active and reactive power, respectively. At node 671, subcontrollers $G_{c1,k}$ and $G_{c2,k}$ have access to the voltage magnitude and angle measurements, where the voltage angles are relative to the slack bus at node 650. At nodes 611 and 634, the active and reactive power is varied to represent an unmodeled load disturbance.

and 632 with unknown multivariable actuator saturation, using decentralized control (Figure 17). The active and reactive power at nodes 634 and 611 is defined using (12) and (13), respectively, and the load disturbance $d(k)$ is

$$d(k) = \begin{cases} 0, & k < 1000, \\ 1, & \text{otherwise.} \end{cases} \tag{27}$$

For node 671, the command is set as

$$r_{671,\text{mag}}(k) = 2267 \text{ V}, \tag{28}$$

$$r_{671,\text{ang}}(k) = -5.3^\circ. \tag{29}$$

To simulate an unknown multivariable-actuator saturation at nodes 632 and 675, on each phase $i \in \{A, B, C\}$, set

$$P_{\text{act,i}} = \begin{cases} P_{\text{act,i,req}}, & P^2_{\text{act,i,req}} + Q^2_{\text{act,i,req}} \le S^2_{\max} \\ S_{\max} \cos\alpha, & \text{otherwise} \end{cases}, \tag{30}$$

$$Q_{\text{act,i}} = \begin{cases} Q_{\text{act,i,req}}, & P^2_{\text{act,i,req}} + Q^2_{\text{act,i,req}} \le S^2_{\max} \\ S_{\max} \sin\alpha, & \text{otherwise} \end{cases}, \tag{31}$$

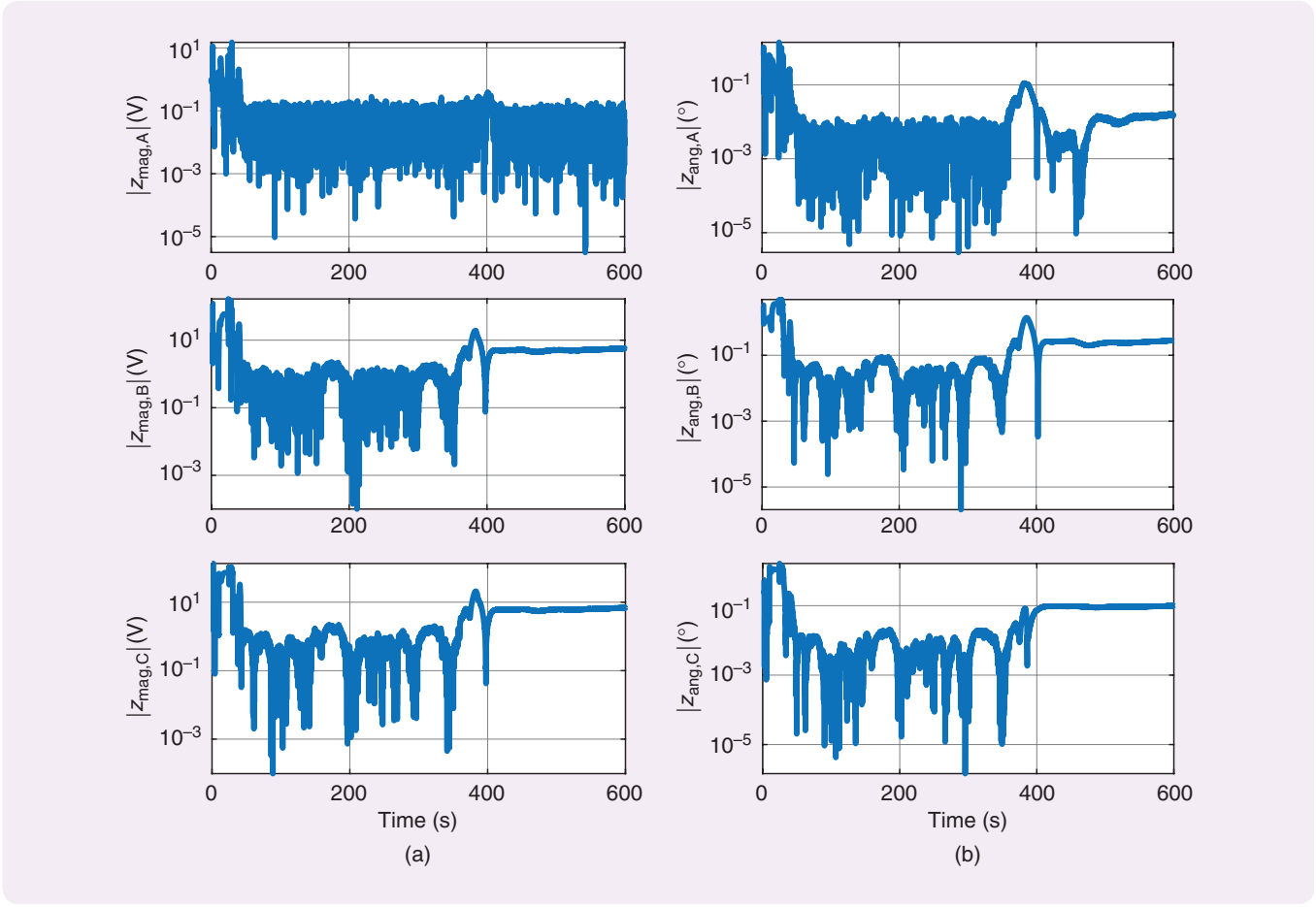


FIGURE 18 Perturbed simulation testing 3: the (a) voltage-magnitude error $|z_{\text{mag}}|$ and (b) voltage angle-magnitude error $|z_{\text{ang}}|$ for node 671 are shown on a logarithmic scale. Asymptotically, the voltage-magnitude and voltage-angle errors are approximately lower than 1 V and 0.1°, respectively.

where

$$\alpha \triangleq \operatorname{atan2}(Q_{\text{act,i,req}}, P_{\text{act,i,req}}). \tag{32}$$

$P_{\text{act,i}}$ and $Q_{\text{act,i}}$ are the active and reactive power extracted at phase i in kilowatts and kvar, respectively; $P_{\text{act,i,req}}$ and $Q_{\text{act,i,req}}$ are the requested active and reactive power extraction at phase i in kilowatts and kvar, respectively; $S_{\max} = 350$ kVA is the maximum apparent power rating for each phase of the inverter; and arguments of k are omitted for clarity. This saturation represents an upper limit on the apparent power extracted or injected at each phase: $S_{\max}$.

For each subcontroller $G_{c1,k}$ and $G_{c2,k}$, the tuning weights described at the beginning of this section are used. Feedback control with adaptation using subcontrollers $G_{c1,k}$ and $G_{c2,k}$ begins at $t = 2$ s and $t = 30$ s, respectively. Figures 18–21 illustrate a pair of decentralized RCAC subcontrollers with voltage magnitude and angle commands at node 671, actuation at nodes 675 and 632, and the time-varying load disturbance (27) at nodes 611 and 634 in the presence of the sensor noise (34). There is no direct communication between subcontrollers $G_{c1,k}$ and $G_{c2,k}$.

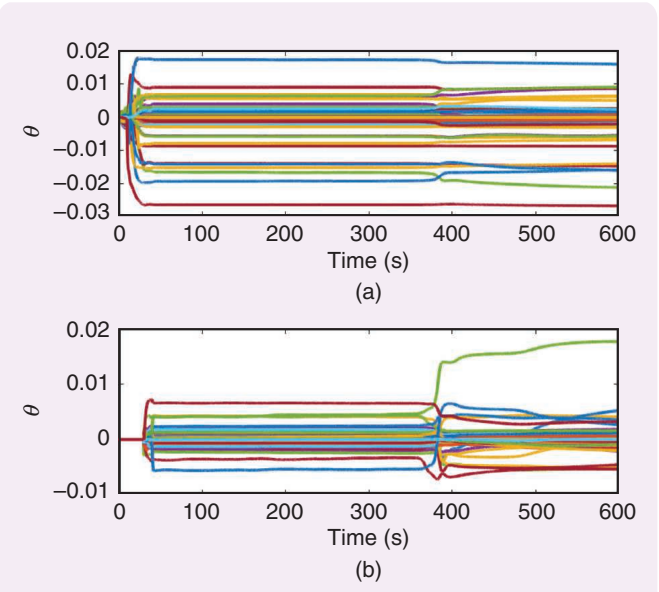


FIGURE 19 Perturbed simulation testing 3: (a) at $t = 2$ s, subcontroller $G_{c1,k}$ begins adapting, and (b) at $t = 30$ s, subcontroller $G_{c2,k}$ begins adapting. The retrospective cost adaptive control has no knowledge of the saturation.

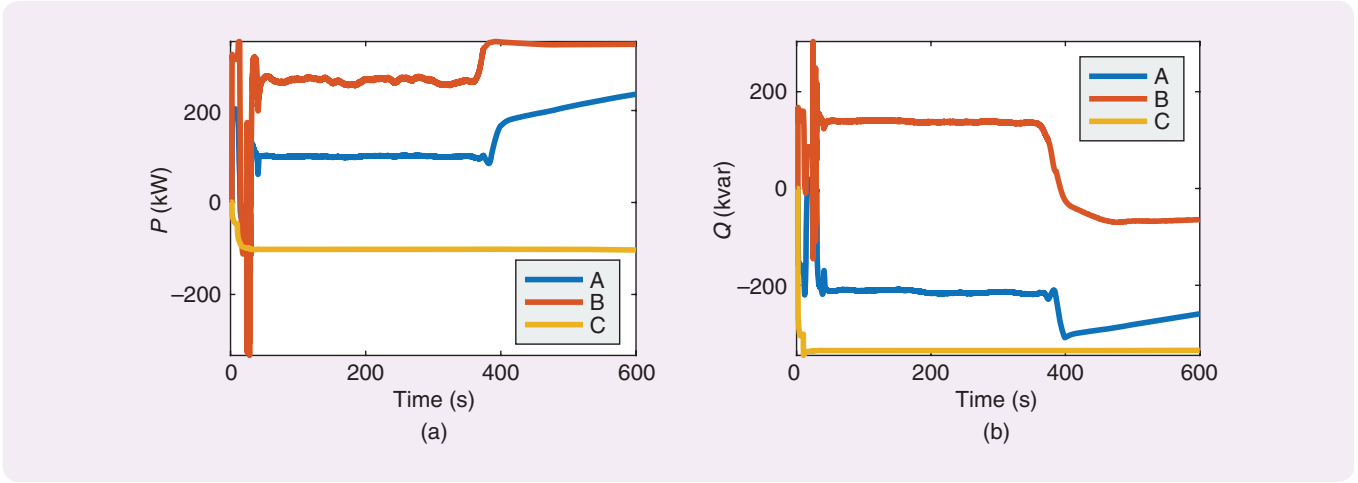


FIGURE 20 Perturbed simulation testing 3: the (a) active and (b) reactive power extracted by subcontroller $G_{c1,k}$ from each phase at node 675 is shown.

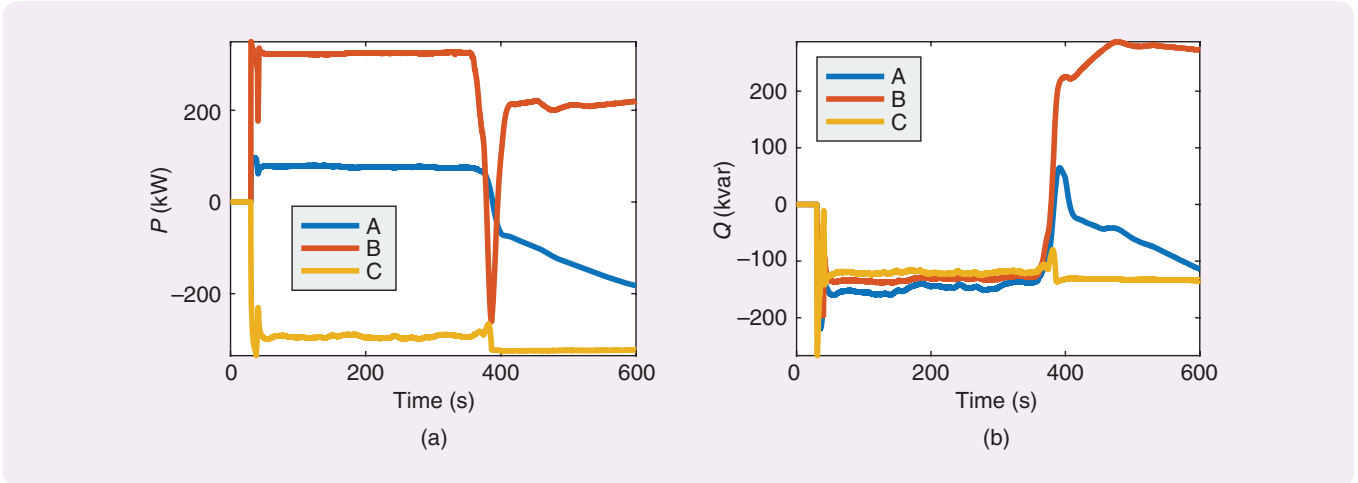


FIGURE 21 Perturbed simulation testing 3: the (a) active and (b) reactive power extracted by subcontroller $G_{c2,k}$ from each phase at node 632 is shown.

In Figure 22, the requested and extracted reactive power is plotted versus the requested and extracted active power for $0 \leq k \leq 60{,}000$ for each phase, A, B, C, at each actuation node. This shows that the unknown multivariable-actuator saturation affects five of the six phases of the two actuation nodes. RCAC has no knowledge of either the saturation form or level. ◊

Perturbed Simulation Testing 4 (PST4)

PST 4 tests command-following at node 671, with PV generation at nodes 611, 634, 645, 646, 652, and 692 and actuation at nodes 675 and 632, using decentralized control (Figure 23). PV penetration in IEEE13NTF is included by using the following procedure. From the Pecan Street online repository [36], 24 h of generation and consumption data (midnight to midnight) are obtained for 200 customers with PV generation capability. Since these data are per minute, linear interpolation is used to obtain subminute data for simulation. Next, for nodes 611, 634, 645, 646, 652, and 692, these data are aggregated using combinations of the 200 customers. The static loads in IEEE13NTF are then replaced at each PV node with the aggregated data. The aggregated data represent a collection of loads with a significant proportion of PV generation. The time-varying load is a disturbance to be rejected by RCAC. For node 671, the command is

$$r_{671,\mathrm{mag}}(k) = 2267 \text{ V}, \tag{33}$$

$$r_{671,\mathrm{ang}}(k) = -5.3°. \tag{34}$$

For each subcontroller $G_{c1,k}$ and $G_{c2,k}$, the tuning weights described at the beginning of this section are used. There is no communication between subcontrollers $G_{c1,k}$ and $G_{c2,k}$. Feedback control with adaptation using subcontrollers $G_{c1,k}$ and $G_{c2,k}$ begins at $t = 2$ s and $t = 30$ s, respectively. Figures 24–27 illustrate a pair of decentralized

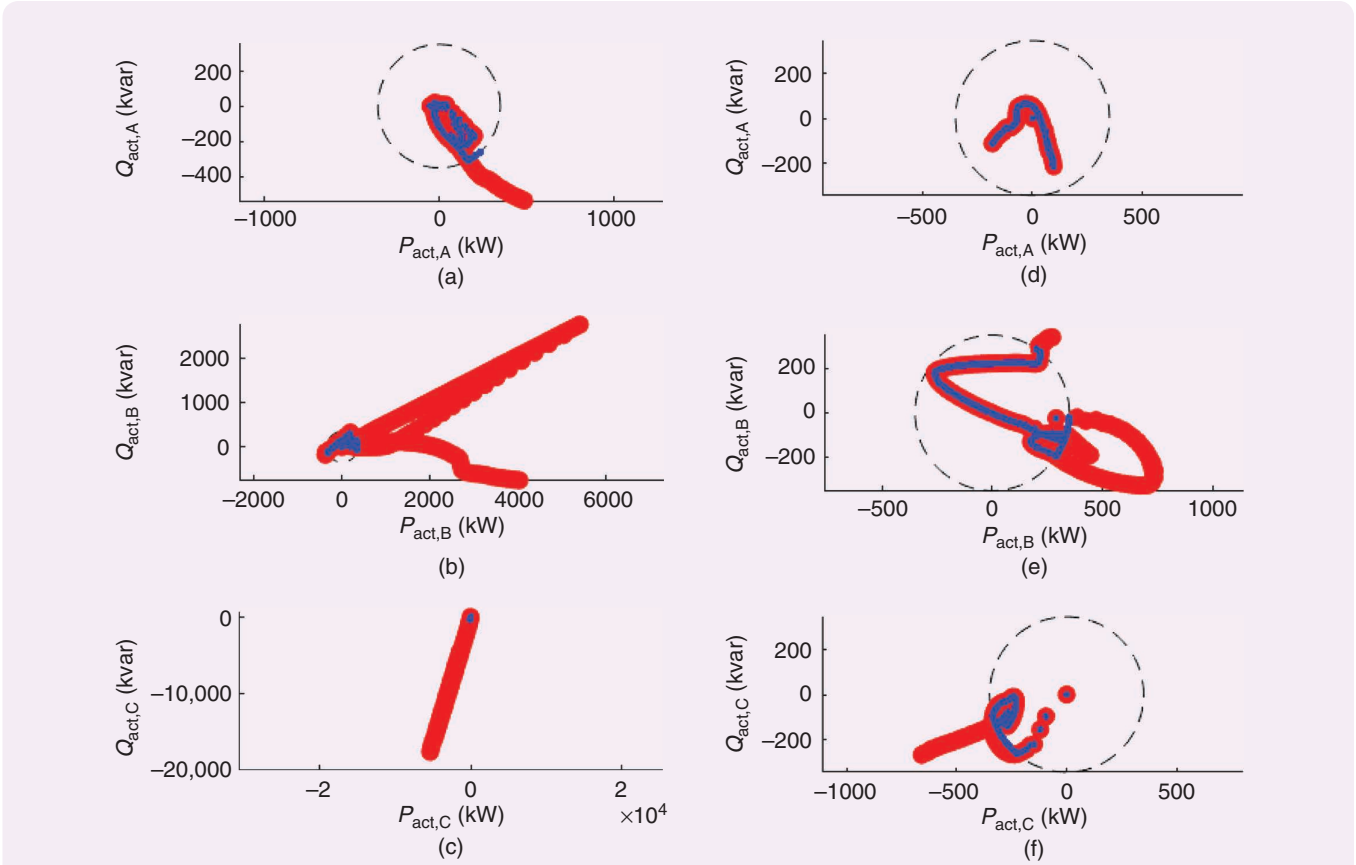


FIGURE 22 Perturbed simulation testing 3: the requested (red) and applied (blue) power of phases A, B, and C at node 675 is shown in (a), (b), and (c), respectively. The requested (red) and applied (blue) power of phases A, B, and C at node 632 is shown in (d), (e), and (f), respectively. The black, dashed circles represent the upper limit of apparent power flow in the actuator S_{max} defined in (30) and (31).

RCAC subcontrollers with voltage magnitude and angle commands at node 671, actuation at nodes 675 and 632, and PV generation at nodes 611, 634, 645, 646, 652, and 692 in the presence of the sensor noise (34). Define

$$\text{PV penetration} \triangleq \frac{\max\limits_t G(t)}{\max\limits_t L(t)} \times 100\%, \tag{35}$$

where $G(t)$ is the total PV generation across all nodes, and $L(t)$ is the total load across all nodes. For this simulation, PV penetration = 70.61%. ♢

CONCLUSIONS

This article provided a numerical investigation of phasor-based control of three-phase power using the IEEE13NTF model provided by the OPAL-RT RT-Lab ePHASORSIM package. Since the details of the ePHASORSIM implementation of IEEE13NTF are inaccessible, the IEEE13NTF model was treated as an opaque executable simulation, and controller tuning was based entirely on simulation. RCAC was applied to this system in two phases: NST and PST.

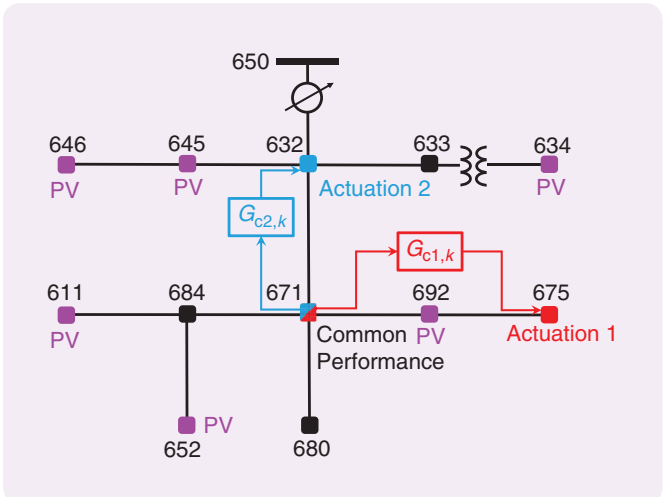


FIGURE 23 Perturbed simulation testing 4: at nodes 675 and 632, subcontrollers $G_{c1,k}$ and $G_{c2,k}$ can add or remove active and reactive power, respectively. At node 671, subcontrollers $G_{c1,k}$ and $G_{c2,k}$ have access to the voltage magnitude and angle measurements, where the voltage angles are relative to the slack bus at node 650. Nodes 611, 634, 645, 646, 652, and 692 emulate diurnal photovoltaic (PV) power generation and consumption to represent unmodeled load disturbances.

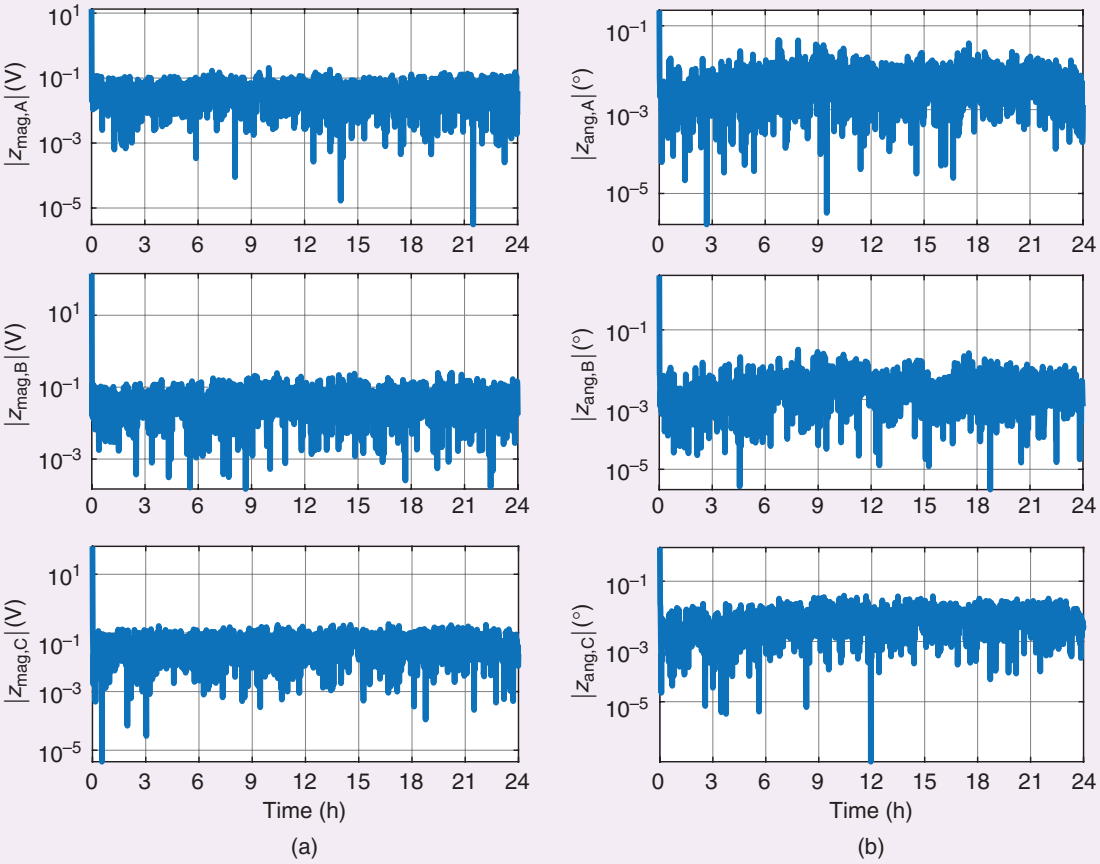


FIGURE 24 Perturbed simulation testing 4: the (a) voltage-magnitude error $|z_{mag}|$ and (b) voltage angle-magnitude error $|z_{ang}|$ for node 671 are shown on a logarithmic scale. Asymptotically, the voltage-magnitude and voltage-angle errors are approximately lower than 0.1 V and 0.01°, respectively.

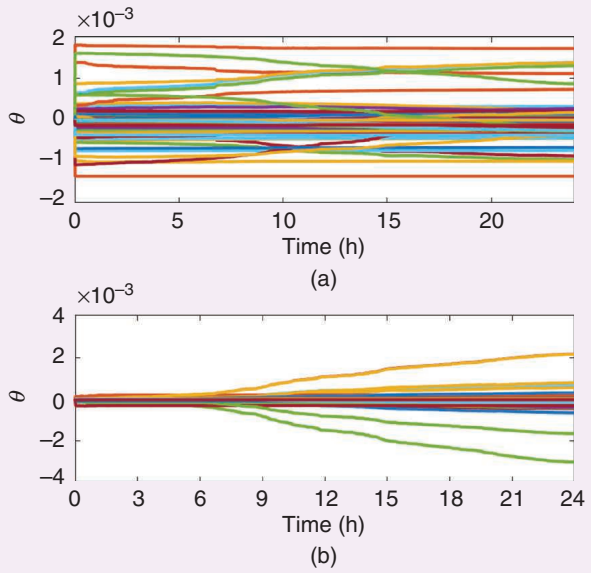


FIGURE 25 Perturbed simulation testing 4: (a) $\theta(k)$ for subcontroller $G_{c1,k}$ and (b) $\theta(k)$ for subcontroller $G_{c2,k}$ are shown.

Based on this approach, the target model used by RCAC was a sparse approximation $\hat{H}_1$ of a single 6×6 impulse-response coefficient H_1. PSTs 1–4 show that $\hat{H}_1$ is robust to changes in the configuration of the actuation and performance nodes. Simulations using the IEEE13NTF model indicate that, with a sparse target model, RCAC provides reliable control with extremely limited modeling information.

To reduce the complexity of the controller for real-time implementation, a fixed-structure extension of RCAC was used, where each SISO entry of each 6×6 MIMO controller was chosen to be either a PI or a second-order IIR controller. RCAC was used in a decentralized architecture, where two subcontrollers were used to follow setpoint commands at physically distinct performance nodes and reject a step-load disturbance at two physically distinct nodes (despite the complete lack of communication between the subcontrollers). For the same scenario, RCAC was shown to be robust to unknown multivariable actuator saturation. Specifically, the two RCAC subcontrollers worked together to achieve a common performance objective despite the limited control authority from each

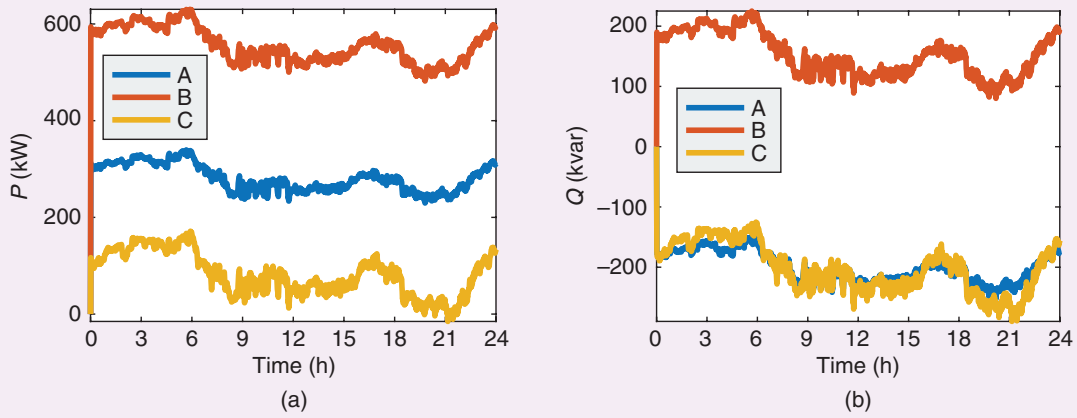


FIGURE 26 Perturbed simulation testing 4: the (a) active and (b) reactive power extracted by subcontroller $G_{c1,k}$ from each phase at node 675 is shown.

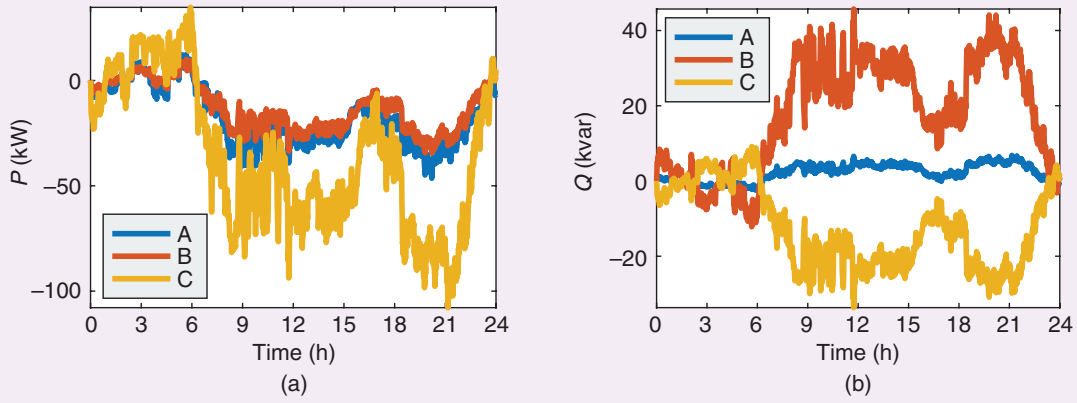


FIGURE 27 Perturbed simulation testing 4: the (a) active and (b) reactive power extracted by subcontroller $G_{c2,k}$ from each phase at node 632 is shown.

controller separately. No adverse interaction between the two controllers was observed. Finally, the ability of two RCAC subcontrollers to follow phasor commands in the presence of 70.61% PV penetration in IEEE13NTF was demonstrated.

Future research will extend IEEE13NTF to include dynamic effects so that it is no longer quasi-static. The goal is to determine the extent of modeling information needed in the presence of unknown dynamics.

ACKNOWLEDGMENTS

The authors are grateful to Alexandra Von Meier for helpful suggestions and discussions. This research was supported by the U.S. Department of Energy under grant DE-EE0008008.

AUTHOR INFORMATION

Syed Aseem Ul Islam (aseemisl@umich.edu) received the B.Sc. degree in aerospace engineering from the Institute of Space Technology, Islamabad, Pakistan, and is currently pursuing the Ph.D. degree in aerospace engineering from the University of Michigan, Ann Arbor. His research interests include data-driven adaptive control for aerospace applications.

Elizabeth L. Ratnam received the B.Eng. (Hons. I) degree in electrical engineering in 2006 and the Ph.D. degree in electrical engineering in 2016 from the University of Newcastle, Australia. She subsequently held postdoctoral research positions with the Center for Energy Research at the University of California, San Diego, and at the University of California, Berkeley, in the California Institute for Energy and Environment. From 2001 to 2012, she held various positions at Ausgrid, a utility that operates one of the largest electricity distribution networks in Australia. She currently holds a Future Engineering Research Leader Fellowship from the Australian National University (ANU), and she joined the Research School of Engineering at ANU as a research fellow and lecturer in 2018. Her research

interests include applications of control theory to power distribution networks.

Ankit Goel received the B.E. degree in mechanical engineering from the Delhi College of Engineering, India, and the Ph.D. degree in aerospace engineering from the University of Michigan, Ann Arbor. His research interests include data-driven estimation and control of high-dimensional complex systems.

Dennis S. Bernstein received the Sc.B. degree in applied mathematics from Brown University, Providence, Rhode Island, and the Ph.D. degree in control engineering from the University of Michigan, Ann Arbor, where he is currently a professor in the Aerospace Engineering Department. His research interests include estimation and control for aerospace applications. He is the author of *Scalar, Vector, and Matrix Mathematics* (Princeton University Press).

REFERENCES

[1] X. Li, D. Hui, and X. Lai, "Battery energy storage station (BESS)-based smoothing control of photovoltaic (PV) and wind power generation fluctuations," *IEEE Trans. Sustain. Energy*, vol. 4, no. 2, pp. 464–473, Apr. 2013.

[2] F. Blaabjerg, R. Teodorescu, M. Liserre, and A. V. Timbus, "Overview of control and grid synchronization for distributed power generation systems," *IEEE Trans. Ind. Elect.*, vol. 53, no. 5, pp. 1398–1409, Oct. 2006.

[3] K. Tanaka et al., "Decentralised control of voltage in distribution systems by distributed generators," *IET Gen. Trans. Distrib.*, vol. 4, no. 11, pp. 1251–1260, 2010.

[4] E. Demirok, D. Sera, R. Teodorescu, P. Rodriguez, and U. Borup, "Evaluation of the voltage support strategies for the low voltage grid connected PV generators," in *Proc. IEEE Energy Conv. Congr. Exposition*, 2010, pp. 710–717.

[5] K. Turitsyn, P. Sulc, S. Backhaus, and M. Chertkov, "Options for control of reactive power by distributed photovoltaic generators," *Proc. IEEE*, vol. 99, no. 6, pp. 1063–1073, 2011.

[6] M. Mills-Price et al., "Solar generation control with time-synchronized phasors," in *Proc. Conf. Protective Relay Engineers*, 2011, pp. 160–167.

[7] M. Savaghebi, A. Jalilian, J. C. Vasquez, and J. M. Guerrero, "Secondary control scheme for voltage unbalance compensation in an islanded droop-controlled microgrid," *IEEE Trans. Smart Grid*, vol. 3, no. 2, pp. 797–807, June 2012.

[8] M. Savaghebi, A. Jalilian, J. C. Vasquez, and J. M. Guerrero, "Autonomous voltage unbalance compensation in an islanded droop-controlled microgrid," *IEEE Trans. Ind. Elect.*, vol. 60, no. 4, pp. 1390–1402, Apr. 2013.

[9] D. E. Olivares et al., "Trends in microgrid control," *IEEE Trans. Smart Grid*, vol. 5, no. 4, pp. 1905–1919, July 2014.

[10] M. A. Abusara, S. M. Sharkh, and J. M. Guerrero, "Improved droop control strategy for grid-connected inverters," *Sustain. Energy Grids Netw.*, vol. 1, pp. 10–19, Mar. 2015.

[11] H. Li, F. Li, Y. Xu, D. T. Rizy, and J. D. Kueck, "Adaptive voltage control with distributed energy resources: Algorithm, theoretical analysis, simulation, and field test verification," *IEEE Trans. Power Syst.*, vol. 25, no. 3, pp. 1638–1647, 2010.

[12] L. B. Perera, G. Ledwich, and A. Ghosh, "Multiple distribution static synchronous compensators for distribution feeder voltage support," *IET Gen. Trans. Distrib.*, vol. 6, no. 4, pp. 285–293, 2012.

[13] A. K. Singh and B. C. Pal, "Decentralized nonlinear control for power systems using normal forms and detailed models," *IEEE Trans. Power Syst.*, vol. 33, no. 2, pp. 1160–1172, Mar. 2018.

[14] D. K. Molzahn et al., "A survey of distributed optimization and control algorithms for electric power systems," *IEEE Trans. Smart Grid*, vol. 8, no. 6, pp. 2941–2962, Nov. 2017.

[15] D. Callaway and I. A. Hiskens, "Achieving controllability of electric loads," *Proc. IEEE*, vol. 99, no. 1, pp. 184–199, Jan. 2011.

[16] W. H. Kersting, "Radial distribution test feeders," in *Proc. IEEE Power and Engineering Society Winter Meeting Conf.*, Jan. 2001, vol. 2, pp. 908–912.

[17] H. Wang and N. N. Schulz, "A revised branch current-based distribution system state estimation algorithm and meter placement impact," *IEEE Trans. Power Syst.*, vol. 19, no. 1, pp. 207–213, Feb. 2004.

[18] F. Jurado, "Modeling SOFC plants on the distribution system using identification algorithms," *J. Power Sources*, vol. 129, no. 2, pp. 205–215, 2004.

[19] S. Moghaddas-Tafreshi and E. Mashhour, "Distributed generation modeling for power flow studies and a three-phase unbalanced power flow solution for radial distribution systems considering distributed generation," *Elec. Power Syst. Res.*, vol. 79, no. 4, pp. 680–686, 2009.

[20] M. Abdel-Akher and K. M. Nor, "Fault analysis of multiphase distribution systems using symmetrical components," *IEEE Trans. Power Del.*, vol. 25, no. 4, pp. 2931–2939, Oct. 2010.

[21] S. Paudyal, C. A. Canizares, and K. Bhattacharya, "Optimal operation of distribution feeders in smart grids," *IEEE Trans. Ind. Electron.*, vol. 58, no. 10, pp. 4495–4503, Oct. 2011.

[22] S. Gautam and S. M. Brahma, "Detection of high impedance fault in power distribution systems using mathematical morphology," *IEEE Trans. Power Syst.*, vol. 28, no. 2, pp. 1226–1234, May 2013.

[23] I. Dzafic, M. Gilles, R. A. Jabr, B. C. Pal, and S. Henselmeyer, "Real time estimation of loads in radial and unsymmetrical three-phase distribution networks," *IEEE Trans. Power Syst.*, vol. 28, no. 4, pp. 4839–4848, Nov. 2013.

[24] H. Yazdanpanahi, Y. W. Li, and W. Xu, "A new control strategy to mitigate the impact of inverter-based DGs on protection system," *IEEE Trans. Smart Grid*, vol. 3, no. 3, pp. 1427–1436, Sept. 2012.

[25] O. Ardakanian, S. Keshav, and C. Rosenberg, "Real-time distributed control for smart electric vehicle chargers: From a static to a dynamic study," *IEEE Trans. Smart Grid*, vol. 5, no. 5, pp. 2295–2305, Sept. 2014.

[26] I. Sharma, K. Bhattacharya, and C. Canizares, "Smart distribution system operations with price-responsive and controllable loads," *IEEE Trans. Smart Grid*, vol. 6, no. 2, pp. 795–807, Mar. 2015.

[27] T. Morstyn, B. Hredzak, and V. G. Agelidis, "Distributed cooperative control of microgrid storage," *IEEE Trans. Power Syst.*, vol. 30, no. 5, pp. 2780–2789, Sept. 2015.

[28] B. Millar, D. Jiang, and M. E. Haque, "Constrained coordinated distributed control of smart grid with asynchronous information exchange," *J. Mod. Power Syst. Clean Energy*, vol. 3, no. 4, pp. 512–525, Dec. 2015.

[29] J. R. Castro, M. Saad, S. Lefebvre, D. Asber, and L. Lenoir, "Optimal voltage control in distribution network in the presence of DGs," *Int. J. Elec. Power Energy Syst.*, vol. 78, pp. 239–247, June 2016.

[30] A. Ghavami, K. Kar, and A. Gupta, "Decentralized charging of plug-in electric vehicles with distribution feeder overload control," *IEEE Trans. Autom. Control*, vol. 61, no. 11, pp. 3527–3532, Nov. 2016.

[31] ePHASORSIM. Accessed on: Oct. 2018. [Online]. Available: https://www.opal-rt.com/systems-ephasorsim/

[32] Y. Rahman, A. Xie, and D. S. Bernstein, "Retrospective cost adaptive control: Pole placement, frequency response, and connections with LQG control," *IEEE Control Syst. Mag.*, vol. 37, pp. 28–69, Oct. 2017.

[33] J. B. Hoagg and D. S. Bernstein, "Retrospective cost model reference adaptive control for nonminimum-phase systems," *J. Guid. Control Dyn.*, vol. 35, no. 6, pp. 1767–1786, 2012.

[34] S. A. Ul Islam and D. S. Bernstein, "Output-feedback adaptive control of discrete-time systems with unknown, unmatched, inaccessible nonlinearities," in *Proc. American Control Conf.*, Milwaukee, WI, June 2018, pp. 1695–1700.

[35] J. Yan and D. S. Bernstein, "Minimal modeling retrospective cost adaptive control of uncertain hammerstein systems using auxiliary nonlinearities," *Int. J. Control*, vol. 87, pp. 483–505, 2014. doi: 10.1080/00207179.2013.842264.

[36] Pecan Street. Accessed on: Oct. 2018. [Online]. Available: http://www.pecanstreet.org/

Environmental Sustainability
https://doi.org/10.1007/s42398-019-00071-3

ORIGINAL ARTICLE

Assessment of bacoside production, total phenol content and antioxidant potential of elicited and non-elicited shoot cultures of *Bacopa monnieri* (L.)

Nupur Jauhari[1] · Rakesh Bharadwaj[2] · Neelam Sharma[2] · Navneeta Bharadvaja[1]

Received: 6 October 2018 / Revised: 3 June 2019 / Accepted: 7 June 2019
© Society for Environmental Sustainability 2019

Abstract
Thirteen accessions of *Bacopa monnieri* (L.) Wettst. were assessed for in vitro propagation, total bacosides, total phenol content and antioxidant potential. Nodal explants were cultured on Murashige and Skoog (MS) medium with 0.4 mg/L benzyl aminopurine (BA) to identify best accession based on growth parameters. The selected accession (IC 554588) was elicited with 1 mg/L of jasmonic acid, salicylic acid or malt extract to assess its efficacy in enhancing secondary metabolites production in in vitro cultures. Elicitor-treated plants showed increased production of bacosides (2.7–3.9 fold), total phenol content (5–18 fold) and higher antioxidant potential (7.9 fold) in 4 weeks. According to statistical analysis, antioxidant potential is highly correlated to total phenol as R-sq (adj) is 90.1%. Data was subjected to correlation, Principal Component Analysis (PCA) and cluster analysis to understand the relationship between different variables and identify major contributors of variability. Strong correlation between total bacoside, total phenol, and antioxidant activity indicate that elevated antioxidant potential was the result of overall increase in bacoside and phenol content, which can be enhanced by the application of elicitors. The study suggested that accessions IC 554588, IC 344312 and IC 554585 are elite and may be utilized for various pharmacopeias after further in vivo research and clinical trials.

Keywords Antioxidant · Bacoside · *Bacopa monnieri* · Elicitor · Phenols · Principal component analysis

Introduction

Secondary metabolites from herbal plants are important source of medicinal compounds. Pharmacological effects of herbal plants has been identified from earliest annals of human habitancy. Therapeutic properties of plants are attributed to secondary metabolites including antioxidant properties (Chen et al. 2009). Antioxidants support human defense system to fight age-related degenerative diseases and other oxidative damages (Atoui et al. 2005). Synthetically produced antioxidants are commonly being used presently. The major drawback of synthetic antioxidants is their adverse side effects (Chen et al. 1992). Although the living creatures have the capacity to balance the fatalistic effect of free radicals, it is necessary to provide some supplements through diet to maintain the radical concentration to lesser extent. Considering the above facts, consumption of herbal drugs rich in antioxidants is an alternate way to strengthen the antioxidant potential of body. Secondary metabolites of plants offer a safe alternative for antioxidants (Walton and Brown 1999).

Bacopa monnieri (L) Wettst. (Family Scrophulariaceae), commonly known as 'Brahmi', 'Neerbrahmi' and 'the thinking person's herb', has been reported to possess antioxidant properties (Kapoor 1990). Whole plant is widely used in *Ayurvedic* formulations since ancient times to enhance memory functions or cognitive functions (Kumar et al. 2016). Brahmi is also utilized as hepatoprotective, cell stabilizer (Sairam et al. 2002), antidepressant, sedative and cardiotonic (Roodernrys et al. 2002). Shahid et al. (2017) suggested the use of *B. monnieri* in neuropathic pain associated with

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s42398-019-00071-3) contains supplementary material, which is available to authorized users.

✉ Navneeta Bharadvaja
navneetab@dtu.co.in

[1] Department of Biotechnology, Delhi Technological University, Delhi, India

[2] ICAR-National Bureau of Plant Genetic Resources, Pusa Campus, New Delhi, India

allodynia and hyperalgesia. These pharmacological effects are attributed to the presence of biologically active secondary metabolites such as brahmins, stigmasterol, beta-sitosterol, triterpenoid saponins, bacosides and bacosaponins (Basu et al. 1967). Researchers have reported the antioxidant property of bacosides (Kapoor 1990; Meena et al. 2012; Alamgir et al. 2014) including free radical scavenging potential (Anbarasi et al. 2005). Bacoside A3 and bacosaponine C acts as inhibitor of superoxide release from the polymorphonuclear cells (Pawar et al. 2001). Madhu et al. (2019) reported the effectiveness of bacoside A against progression of experimental autoimmune encephalomyelitis by reducing neutrophil infiltration and demyelination. The requirement of medically useful secondary metabolites bacosides has increased significantly during last two decades. Demand of this key active herbal principle is increasing continuously as researches have reported its potential to cure various neurodegenerative diseases and improve cognitive functions. Estimated annual trade of brahmi is 2000–5000 MT with the price range of Rs. 30–35 per kg (NMPB 2019). Unorganized collection and over-exploitation of plant from its natural habitat are responsible for threat to its existence (Verma et al. 2012). Therefore, there is an urgent need to manage important medicinal plants for present and future generations. To facilitate sustainable use, germplasm collection and use of different accessions for plant tissue culture technique is essential. Plant tissue culture technique has tremendous potential for sustainable use of overexploited medicinal plants like *B. monnieri*. Plant tissue culture based propagation methods play an important role in the production of secondary metabolites (Ahmad et al. 2013). Higher bacoside A content was reported in in vitro cultures of *B. monnieri* as compared to field grown plants (Sharma et al. 2012). The yield of secondary metabolites is strongly affected by environmental conditions (Gorelick et al. 2014). According to literature, elicitors increase the yield of secondary metabolites (Hussain et al. 2012; Karuppusamy, 2009). Although, several abiotic chemical and plant hormones such as jasmonic acid (JA) (Farmer et al. 2003; Sharma et al. 2015), salicylic Acid (SA) (Ali et al. 2006; Sharma et al. 2015), malt extract (ME) (Kundu et al. 2016) and CuSO4 (Sharma et al. 2015) have been used as elicitors to enhance the secondary metabolite production, their effects are species specific. JA and SA are hormonal elictors involved in the stress response and affect the production of secondary metabolites like terpenoids, flavonoids and alkaloids (Farmer et al. 2003; Shabani et al. 2009). ME can be added as natural mixture in special cases to support culture growth (Mangathayaru 2013).

High therapeutic value of the plant and less availability for future use, justify the sustainable use of *B. monnieri.* Therefore, investigating different accessions of *B. monnieri* in order to select accessions with higher bacoside content and strategy to further enhance bacosides for commercial exploitation is desirable.The present study aims to find elite accession with maximum number of shoots, shoot length, maximum bacosides production and highest antioxidant potential. In addition, elicitation of the selected accession with JA, SA or ME for the assessment of their effect on total bacosides, total phenol content and antioxidant potential compared to non-elicited plants has been evaluated in order to achieve high amount of therapeutic principle.

Materials and methods

Plant material

Thirteen accessions of *B. monnieri* procured from ICAR-National Bureau of Plant Genetic Resources (ICAR-NBPGR) and maintained in tissue culture laboratory, at Delhi Technological University, Delhi, India were used in the present study. All the accessions belong to different states of India (Table 1).

In vitro shoot induction and elicitation

In vitro cultures were maintained on Murashige and Skoog (MS) (Murashige and Skoog 1962) medium supplemented with 0.5 mg/l benzyl aminopurine (BA) (Sharma et al. 2007). For elicitation study, stock solutions were prepared in 90% ethanol for JA and sterile Mili-Q water (Mili-Q, USA) for SA and ME (Sigma-Aldrich USA). Filter sterilized 1.0 mg/l of JA, SA or ME were added to the autoclaved medium. Stem segments 1.0-2.0 cm (2-3 nodes) of selected accession from in vitro propagated plants were transferred to Erlenmeyer flasks (100 mL) containing MS + 0.5 mg/l BA (20 mL) with or without elicitor (control). The cultures were maintained at 25 ± 2 °C under 16-h photoperiod. After 4 weeks the elicitor-treated and non-elicited cultures were evaluated for growth parameters (number of shoots/explant and shoot length), moisture content, antioxidant potential and biochemical parameters for the presence and quantity of major secondary metabolites (total bacosides and total phenol content). All treatments were performed in triplicate and each treatment consisted of 10 samples.

Biochemical analyses

For estimation of total bacosides, total phenol content and antioxidant potential, elicited and non-elicited in vitro cultured *B. monnieri* accessions were analyzed. Cultures from the following treatments were used for sample preparation: (i) non-elicited 4-week-old cultures of 13 accessions grown on MS + 0.5 BA and (ii) 4-week-old cultures of selected

Table 1 List of *B. monnieri* accessions used for the study

Number representation	Accession no.	State of collection	Alphabetical representation	Latitude and longitudee	Temperature (°C)	Annual rainfall (mm)
1	IC 439118	Jharkhand	A	23.610°N 85.279°E	26	1397
2	IC 426442	Punjab	B	31.147 75.341	24	200
3	IC 426447	Madhya Pradesh	C	22.974 78.656	32	1017
4	IC 468878	Kerala	D	10.850 76.271	32	3055
5	IC 373640	Kerala	E	10.850 76.271	32	3055
6	IC 344312	Karnataka	F	15.317 75.713	27	1248
7	IC 531621	Jharkhand	G	23.612 85.279	24	1397
8	IC 375976	Jammu & Kashmir	H	34.083 74.797	21	1238
9	IC 353203	New Delhi	I	28.704 77.102	28	800
10	IC 554588	Madhya Pradesh	J	22.974 78.656	32	1017
11	IC 554586	Madhya Pradesh	K	22.974 78.656	32	1017
12	IC 554587	Madhya Pradesh	L	22.974 78.656	32	1017
13	IC 554585	Amarkantak	M	22.682 81.753	24	800

accession grown on MS + 0.5 BA supplemented with elicitors and control (without elicitor).

Sample preparation

Approximately 5 g of green shoots were harvested, shade-dried and crushed to fine powder. The powdered sample was extracted with HPLC grade methanol (Merck Chemicals) using Soxhlet apparatus (5 cycles) at room temperature. The extracted solvent was dried on rotary evaporator and volume was made up to 2 mL. The sample was filtered through 0.22-micron filter and used for estimation of total bacosides, total phenol content and antioxidant potential by 1,1-diphenyl-2-picrylhydrazyl (DPPH) and cupric ion reducing antioxidant capacity (CUPRAC) assays.

Quantification of total bacosides

Total bacosides were quantified through high performance liquid chromatogrpahy (HPLC) using the method described by Murthy et al. (2006) and Mishra et al. (2013). HPLC experiments were performed on Agilent HPLC system equipped with 1260 Infinity Binary Pump (G1312B). ZORBAX ODS C18 column (5 µm 4.6 × 250 mm), 1260 infinity diode array detector (G4212B). The mobile phase consisted of sodium acetate buffer and acetonitrile 0.1 M (65:35 v/v), pH 3.2 was adjusted using acetic acid. The flow rate was 0.5 mL/min for isocratic elution of bacosides with 30 min run time. Detection was done at 205 nm and sample injection volume was kept at 2.5 µL. Bacoside mixture (1 mg) (Sigma-Aldrich) comprising of bacoside A, bacoside A3, bacoside II, bacoside X and bacosaponin C was dissolved in 2 mL methanol and used as a standard solution to quantify total bacosides.

Linearity and precision

Standard curve was drawn by injecting different concentrations of bacoside. Linear standard curve with $R^2 > 0.9$ was obtained. Good precision based on three replications was observed. Standard deviation within the reasonable level was noticed.

Limit of detection and limit of quantification

Limit of detection and limit of quantification were calculated 0.0306 µg/5 mL and 0.0918 µg/5 mL respectively.

Determination of total phenol content (TPC)

TPC was estimated by Folin-Ciocalteu (FC) reagent using gallic acid as standard solution (Singleton et al. 1999) with slight modifications. 100 µL of the extract was added to 900 µL of distilled water and 200 µL of 1 N FC reagent and 2.0 mL of sodium carbonate (7% w/v). Contents were mixed and allowed to stand for 30 min at room temperature (25 ± 1 °C) in dark. Absorbance was measured at 750 nm using UV–Vis spectrophotometer (LAMBDA 25 UV/Vis Systems, PerkinElmer, Inc. MA 02451, USA). Total phenol content was expressed as µg/g gallic acid equivalent (GAE).

Determination of antioxidant potential

Antioxidant potential of samples was assessed using DPPH and CUPRAC assay.

DPPH assay was done according to Brand-Williams et al. (1995) with slight modification. DPPH assay was performed by pipetting out 800 µL (100 mM) of Tris HCl buffer (pH 7.4) to which 200 µL of standard solution (gallic acid) of varying concentration was added. Different volumes of sample extract were dried under nitrogen and redissolved in 200 µL of methanol, which was added to 0.8 mL of DPPH reagent. The reaction mixture was vortexed and left for 30 min in dark. Absorbance was recorded at 517 nm using UV–vis spectrophotometer (LAMBDA 25 UV/Vis Systems, PerkinElmer, Inc. MA 02451, USA) with methanol as the blank.

$$50\%\text{Inhibition} = \{A\,(\text{control}) - A\,(\text{sample})/A\,(\text{control})\} \times 100$$

where, *A* (control) is absorbance of blank control (containing all reagents except the extract) and *A* (sample) is absorbance of the test sample.

For CUPRAC assay method of Apak et al. (2004) was used. In reaction mixture 1 mL $CuCl_2$ solution (0.01 M), 1 mL neocuproine alcoholic solution (7.5 mM), 1 mL ammonium acetate buffer (1.0 M) of pH 7.0, 100 µL extract and 1 mL water were added (total volume, 4.1 mL). Absorbance against reagent blank was measured at 450 nm after 30 min.

Statistical analysis

Data for in vitro propagation and elicitation was analyzed using ANOVA with SPSS version 21.0 statistical software package. Results were expressed in terms of means and standard error, while means were compared with Duncan's multiple range test (DMRT, 9) at $P < 0.05$ level. Analysis of variance was carried out using PROC GLM to determine significant differences in parameters studied among *B. monnieri* accessions. Multilinear regression (MLR) has been used for statistical analysis to check the correlation among total bacoside, total phenol and antioxidant potential.

Simple linear correlation analysis was performed to indicate the measure of correlation and strength of relationship between variables using Pearson's correlation. Sample similarities were calculated on the basis of pair-wise Euclidean distance and unweighted pair-group method with arithmetic averaging (UPGMA) algorithm was used for establishing cluster to search natural groupings among accessions for different parameters. Principal component analysis (PCA) using correlation matrix was carried out to evaluate relative contribution of moisture, CUPRAC, DPPH, total bacosides and total phenols to total variability in *B. monnieri* accessions.

Result and discussion

In vitro shoot induction and elicitation

Shoot induction and proliferation were assessed for thirteen accessions of *B. monnieri* on MS medium supplemented with 0.5 mg/L BA using nodal explants. Various researchers have reported the growth of *B. monnieri* on different media (Sharma et al. 2007; Jauhari et al. 2016). In the present study, maximum shoot length was observed in IC554588 (2.8 cm) followed by IC375976 (2.2 cm) whereas maximum number of shoots/explant were recorded in IC344314 (8.7 shoots/explants) followed by IC554588 (8.3 shoots/explants) (Table 2). Overall, maximum shoot growth was observed in IC554588 which was selected for elicitation experiments. No alteration in the morphology of the plant and no callus formation were observed in elicitor-supplemented medium. Leonard et al. (2018) reported that inoculum size, sucrose concentration and concentration of KH_2PO_4 are some potent factors for maximum biomass production. Slight reduction in number of shoots/explants and shoot length was observed in elicitor-treated plants (Figs. 1 and 2). This may be due to increase in the level of enzyme β-amyrin synthase, responsible for increasing structural components, up to 2–3 weeks of elicitation with methyl jasmonate. Subsequent reduction of the enzyme after 4 weeks results in reduced number of shoots/explants and shoots length (Mangas et al. 2009). Similar elicitation effect has been

Table 2 Total bacoside amount, total phenol content and antioxidant potential (DPPH and CUPRAC) studied in 13 accessions of *B. monnieri* (culture period—4 weeks)

Accession No.	Moisture (%)	CUPRAC (µg/g)	DPPH (µg/g)	Bacosides (µg/g DW)	Phenol (µg/g)	No of shoots/explant	Shoot length (cm)
IC 439118	86.0 ± 0.5^{e}	60.7 ± 0.4^{j}	0.978 ± 0.6^{g}	324.1 ± 0.6^{h}	17.7 ± 0.6^{b}	7.0 ± 0.8^{cde}	2.1 ± 0.6^{ab}
IC 426442	89.0 ± 0.4^{b}	67.8 ± 0.6^{i}	0.856 ± 0.4^{g}	250.9 ± 0.5^{k}	9.81 ± 0.7^{e}	8.1 ± 0.7^{abc}	2.1 ± 0.6^{ab}
IC 426447	87.5 ± 0.5^{d}	69.9 ± 0.2^{h}	0.681 ± 0.4^{g}	434.0 ± 0.7^{c}	7.58 ± 0.4^{f}	5.6 ± 0.8^{fg}	1.2 ± 0.7^{bc}
IC 468878	88.4 ± 0.7^{cb}	35.1 ± 0.4^{l}	2.32 ± 0.6^{f}	433.4 ± 0.4^{c}	1.49 ± 0.5^{i}	4.5 ± 0.6^{g}	1.1 ± 0.5^{bc}
IC 373640	88.4 ± 0.4^{cb}	51.7 ± 0.9^{k}	3.82 ± 0.5^{e}	407.2 ± 0.5^{f}	5.57 ± 0.5^{g}	6.4 ± 0.6^{def}	1.3 ± 0.6^{bc}
IC 344312	90.1 ± 0.6^{a}	193.0 ± 0.6^{b}	6.49 ± 0.6^{b}	455.1 ± 0.6^{b}	18.3 ± 0.6^{b}	8.7 ± 0.6^{a}	2.1 ± 0.4^{ab}
IC 531621	87.6 ± 0.3^{dc}	88.9 ± 0.4^{f}	3.79 ± 0.6^{e}	328.5 ± 0.7^{g}	5.77 ± 0.7^{g}	5.4 ± 0.7^{fg}	2.1 ± 0.5^{ab}
IC 375976	87.4 ± 0.4^{d}	126.1 ± 0.5^{d}	5.26 ± 0.4^{c}	313.5 ± 0.9^{j}	3.31 ± 0.7^{h}	5.3 ± 0.5^{fg}	2.2 ± 0.7^{ab}
IC353203	89.1 ± 0.5^{b}	108.2 ± 0.4^{e}	4.37 ± 0.9^{e}	208.9 ± 0.1^{l}	15.3 ± 0.4^{c}	7.2 ± 0.6^{bcd}	1.6 ± 0.4^{bc}
IC 554588	88.0 ± 0.6^{dc}	173.4 ± 0.7^{c}	7.15 ± 0.5^{a}	515.4 ± 0.6^{a}	11.9 ± 0.5^{d}	8.3 ± 0.7^{ab}	2.8 ± 0.5^{a}
IC 554586	82.7 ± 0.5^{g}	126.4 ± 0.5^{d}	3.62 ± 0.5^{e}	432.2 ± 0.6^{d}	8.98 ± 0.8^{e}	2.2 ± 0.5^{h}	0.8 ± 0.3^{c}
IC 554587	84.7 ± 0.3^{f}	78.1 ± 0.6^{g}	2.07 ± 0.6^{f}	412.9 ± 0.5^{e}	16.2 ± 0.5^{c}	4.9 ± 0.6^{g}	1.2 ± 0.5^{bc}
IC 554585	85.5 ± 0.6^{e}	231.0 ± 0.5^{a}	4.25 ± 0.4^{d}	318.7 ± 0.4^{i}	22.3 ± 0.6^{a}	5.8 ± 0.5^{efg}	1.5 ± 0.4^{bc}

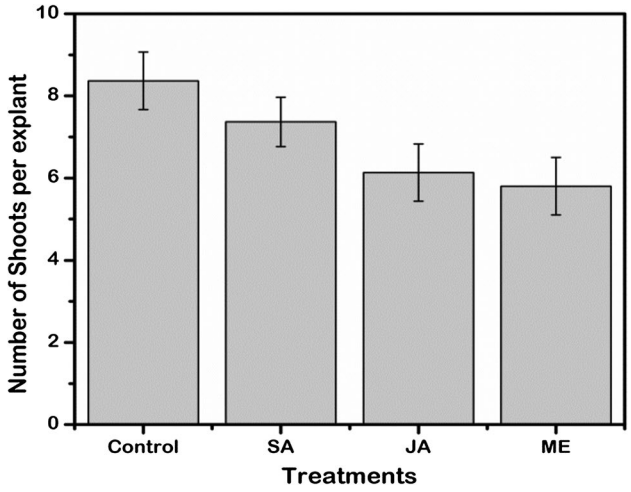


Fig. 1 Effect of elicitors JA, SA and ME (1.0 mg/L) on number of shoots/explant of *B. monnieri* (IC 554588) after 4 weeks of culture. Data represents mean ± standard error of three replicates; each experiment was repeated thrice ($P < 0.05$)

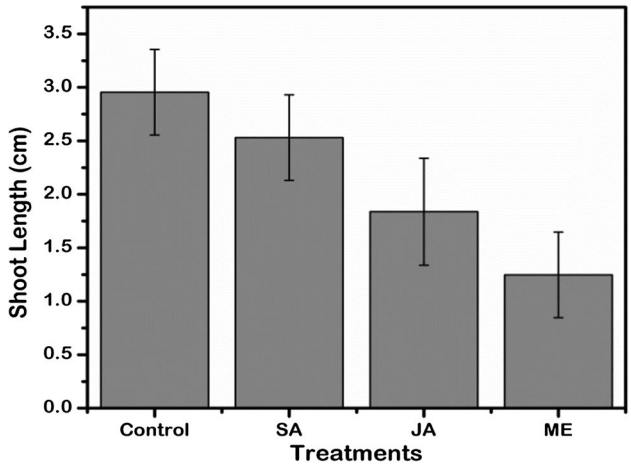


Fig. 2 Effect of elicitors JA, SA and ME (1.0 mg/L) on shoot length (cm) of *B. monnieri* (IC 554588) after 4 weeks of culture. Data represents mean ± standard error of three replicates; each experiment was repeated thrice ($P < 0.05$)

reported in earlier studies for bacoside A and triterpenoid saponins (Namdeo 2007; Sharma et al. 2015). Explants grown on media with ME often displayed necrosis of meristametic area. Thus, the growth of ME-treated plants was inhibited and exhibited less number of shoots/explants and shoot length compared to SA and JA. Plants elicited with SA, exhibited higher number of shoots/explant compared to JA and ME. Shoot length of SA-treated plants was comparable to control while slight decrease was observed with JA and ME. Maximum number of shoot regeneration has been reported in *Centella asiatica* elicited with ME as compared to JA and SA (Kundu et al. 2016). Thus, the response to elicitors may be species specific. Moisture content was more or less constant across all the accessions and only non-significant differences were observed in non-elicited plants (Table 2).

Quantification of total bacosides

Bacoside content in in vitro propagated shoots of elicited and non-elicited plants using HPLC is depicted in Table 2. Wide variation in bacoside amount (216.93–515.47 µg/g dry weight (DW) basis) was observed in 13 accessions based on DMRT. Although accession IC554588 with maximum

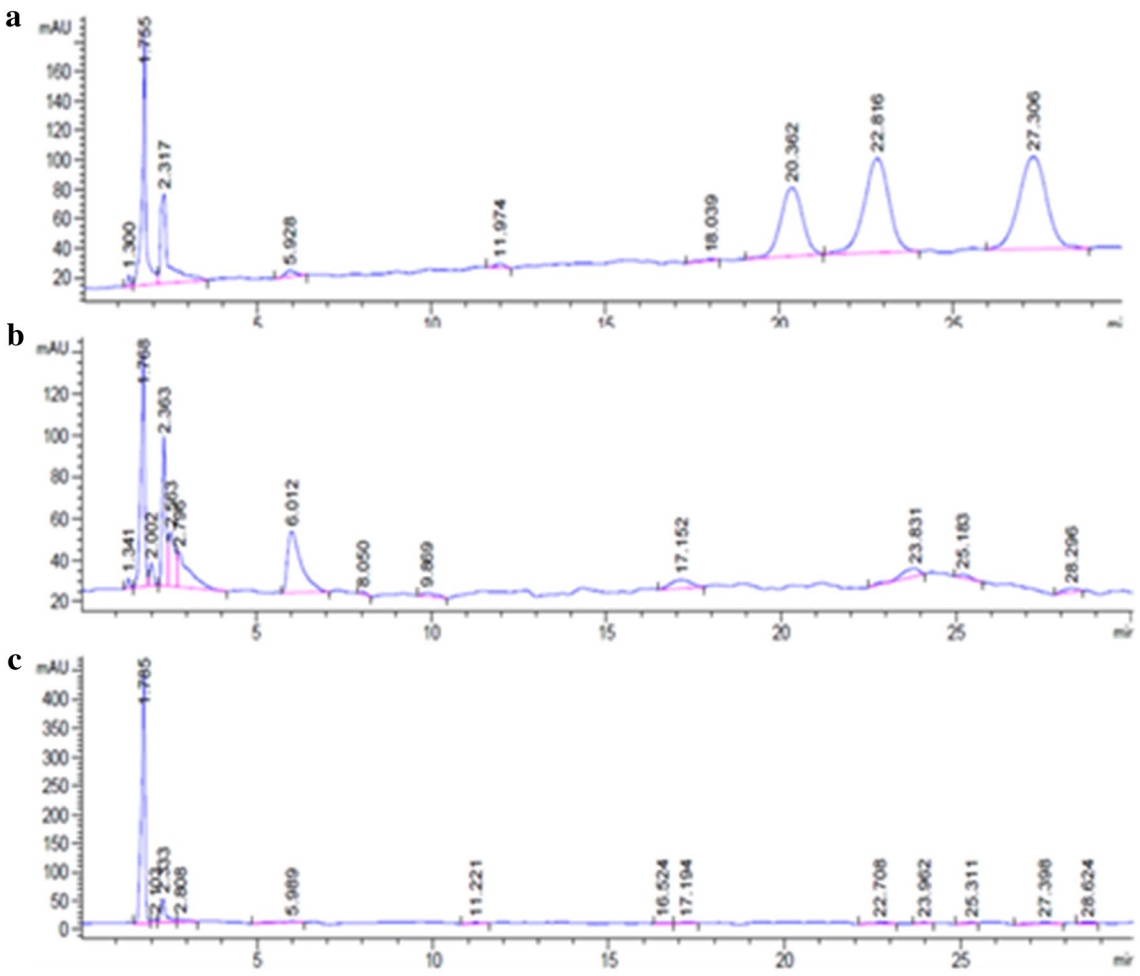


Fig. 3 Chromatogram of bacosides obtained through HPLC **a** standard, **b** control, **c** elicitor treated (JA) for the accession IC554588

number of shoots/explants and highest shoot length produced maximum amount of total bacosides, there was no correlation of growth parameters on yield of bacosides. Accessions belonging to Madhya Pradesh (IC 554588) and Karnataka (IC 344312) produced highest amount of bacosides followed by Kerala (IC 648878), Madhya Pradesh (IC 426447, IC 554586) and Jharkhand (IC 531621, IC 439118). The results of present study are in accordance with earlier reports (Mishra et al. 2013; Mundkinajeddu et al. 2005; Pal et al. 1998). In the present study, IC554588 produced maximum amount of total bacosides (515.4 µg/g DW), which was remarkably influenced by elicitors. Bacoside standard is a mixture of five types of bacosides (bacoside A, bacoside A3, bacoside II, bacopaside X and bacosaponin C) shown at Retention Time 1.7, 2.3, 20.3, 22.8, 27.3 min in chromatogram (Fig. 3a). After analyzing the peaks of 13 accessions, few peaks were below Limit of Quantification (LOQ) but above Limit of Detection (LOD). Out of five peaks of accession IC 426447, four can be quantified while one was only detected. Likewise, in accessions IC 375976 and IC 353203, two out of three peaks and three out of five peaks were quantified respectively.

Elicitors (JA, SA and ME) analyzed in the present investigation, exhibited different effects on bacoside production. Five peaks were observed in bacoside standard as given in Fig. 3a. Two types of bacosides are present in control at Retention Time 1.7, 2.3 min (Fig. 3b). After elicitation of JA, increased amount of bacoside was recorded at Retention Time 1.7, 2.3, 22.8, 27.3 min (Fig. 3c). Despite slight reduction in growth parameters, increased bacoside content was recorded in all elicited samples of accession IC554588 compared to control (Fig. 4a). Maximum amount of total bacosides was obtained in plant elicited with JA (3.9 fold) followed by SA (3.2 fold) and ME (2.7 fold) after 4 weeks of elicitation. Sharma et al. (2015) reported 3.08 and 1.32 fold increase in bacoside content after elicitation with JA and SA respectively in *B. monnieri* after 9 days of elicitation. A 2–3 fold enhancement in bacoside A content was achieved with MeJ (50 µM) and SA (50 µM) respectively, after 3 weeks of elicitation by

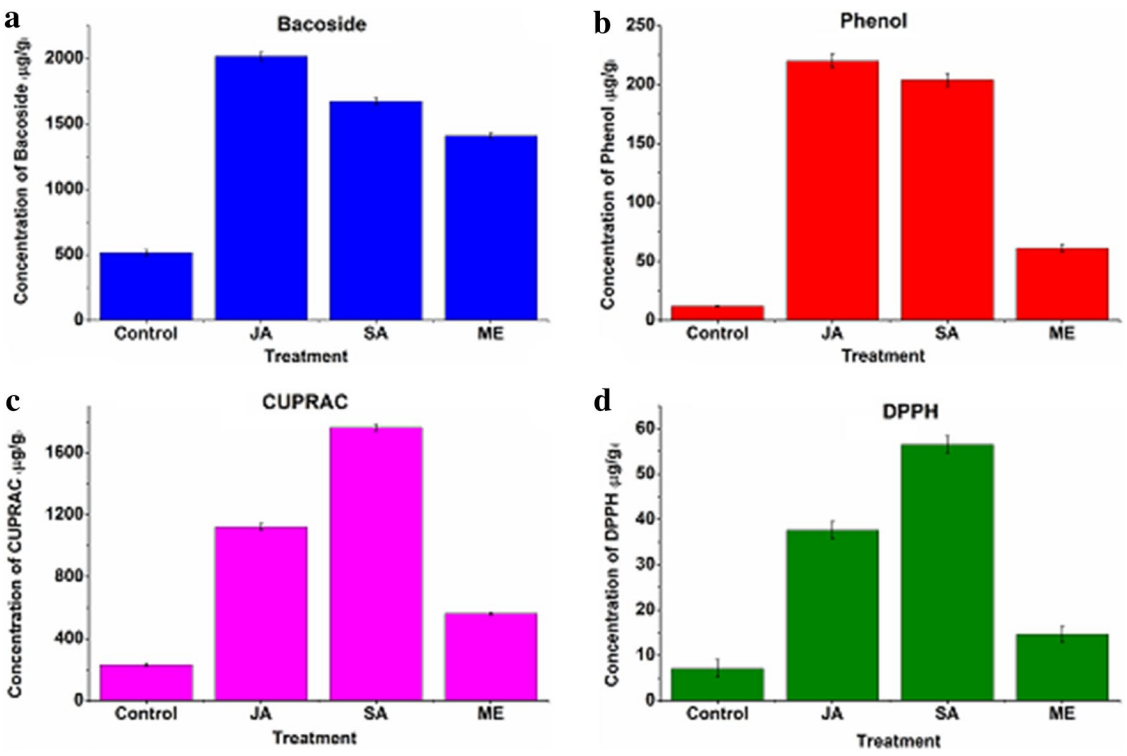


Fig. 4 Effect of elicitors JA, SA and ME (1.0 mg/L) on total bacoside content, total phenol content, CUPRAC and DPPH of *B. monnieri* (IC 554588) after 4 weeks of culture. Data represents mean ± standard error of three replicates; each experiment was repeated thrice ($P < 0.05$)

Largia et al. (2015). 2374 unique transcripts (UTs) were noticed to be signaling up-/down regulated with 24 h elicitation of MeJA in *Polygonum minus* leaves. These UTs contains many genes related to plant secondary metabolites (Rahnamaie-Tajadod et al. 2017). Saeed et al. (2017) reported enhanced levels of phytoecdysteroids with 14 days of methyl jasmonate treatment in *Ajuga bracteosa.* Lee et al. (2004) reported that treatment with methyl jasmonate over express the gene *PgSS1,* which is responsible for upregulation of triterpene saponins. This may be the possible reason for enhanced production of bacoside in JA treated roots. The results of present study are in conformity of earlier findings and exhibit significant elicitation on bacoside content. Sharma et al. (2013) analyzed shoots (in vitro) for bacoside A and reported 4.4 mg/g dry weight of bacoside A and 1.8 fold increase with 50 µM MJ treatment for 1 week. In the present study, high amount of bacoside i.e. 515.4 µg/g DW was obtained in accession IC554588 whereas different accessions showed varied bacoside content (Table 2). The yield of bacoside depends on various factors such as extraction method, media composition on which in vitro plants are grown, age of plants used for the extraction of bacosides and environmental conditions under which the experiment has been performed. In previous reports researchers have analyzed yield of bacoside A only while in our study total bacoside content (bacoside A, bacoside A3, bacoside II, bacoside X and bacosaponin C) was analyzed. Methods for analyzing the amount of bacoside play important role in its yield; many researchers have used high performance thin-layer chromatography (HPTLC) while HPLC was used in present investigation. A 1.8 fold increase in the amount of bacoside A was obtained with treatment of 50 µM MJ for 1 week whereas higher amount of bacoside (3.9 fold) with less amount of elicitor (1 mg/L) but for comparatively long duration (4 weeks) was noticed in the present investigation. Sharma et al. (2015) observed increase in bacoside content after 1 and 4 weeks of elicitor treatment with copper sulfate, JA and SA, whereas, increase in bacoside content was analyzed after 4 weeks of elicitor treatment with ME in place of copper sulphate in the present study. Therefore, variation in experimental conditions in different experiments play major role in the yield of bacosides. Thus, it can be summarized that difference in yield of bacoside and increase in amount of bacoside after elicitation depends on method of metabolite extraction and experimental conditions, which are responsible for variation in bacoside content among accessions.

Determination of total phenol content (TPC)

Bacopa accessions were investigated for TPC, responsible for antioxidant property (Gottlieb and Borin 2000). Phenol plays an important role in free radical scavenging potential of plants. The present study reported 1.5–22.4 µg/g GAE TPC in thirteen accessions of *B. monnieri* (Table 2). Maximum phenol content was noticed in accession belonging to Amarkantak (IC 554585) and Karnataka (IC 344312). Jharkhand (IC 439118) Madhya Pradesh (IC 554587), New Delhi (IC 353203) and other states showed comparatively less amount of phenol content. Elicitors were applied on selected accession IC554588 having 11.9 µg/g GAE phenol content, which was further increased to 219.9, 203.2 and 61.2 µg/g GAE in plant cultures elicited with JA, SA and ME respectively (Fig. 4b). Results of present investigation confirmed earlier reports of phenol concentration in ethanolic and aqueous extract of *B. monnieri*, which was 3.18 µg/mL and 3.71 µg/mL GAE respectively (Mukherjee et al. 2011). Saeed et al. (2017) reported enhanced TPC with the use of methyl jasmonate and phenyl acetic acid in root suspension of *A. bracteosa*. Mendoza et al. (2018) reported that elicitation with 3 µM MeJA after 96 h of treatment produced highest amount of phenol compounds 49.25% higher than control in *Thevetia peruviana*. The present study clearly indicates that elicitation increases accumulation of total phenol content in in vitro cultures of *B. monnieri*.

Antioxidant potential

In the present study, presence of significant amount of bacosides and phenol in the samples of *B. monnieri* suggested that these might be the contributors to antioxidant activity, which can be enhanced with the use of elicitors. Evaluation of antioxidant potential of the plants using different methods (DPPH, CUPRAC, 2,2'-azino-bis-3-ethylbenzthiazoline-6-sulphonic acid (ABTS), ferric reducing antioxidant power (FRAP)) show variation in antioxidant activity (Thaipong et al. 2006). Different methods vary in terms of principle and environmental conditions. It is difficult to precisely measure the antioxidant activity of a system having various components using single test. Therefore, different mechanisms are needed to confirm antioxidant potential of the plants. In the present study, antioxidant potential was studied using DPPH and CUPRAC method. Maximum antioxidant activity was 7.15 µg/g GAE (IC 554588) for DPPH and 231 µg/g GAE (IC 554588) for CUPRAC, among all the thirteen accessions of *B. monnieri*. Antioxidant potential (CUPRAC and DPPH) of all the samples of *B. monnieri* is given in Table 2. Considering both the methods (CUPRAC and DPPH), Madhya Pradesh (IC 554588), Karnataka (IC 344312) and Amarkantak (IC 554585) accessions exhibited maximum antioxidant potential compared to other states among the thirteen accessions of *B. monnieri*. Jain et al. (2017) studied free radical scavenging activity with CUPRAC assay resulting 0.226 TE mM at 1000 µg/mL. Few researchers have evaluated effect of bacosides on antioxidant potential. Hazra et al. (2017) reported stimulated reactive oxidation species (ROS) production and plastid terminal oxidase (PTOX) accumulation by increasing their mRNA stability with the treatment of methyl jasmonate in *Podophyllum hexandrum*. Tripathi et al. (1996) compared the alcoholic extract of *B. monnieri* with Tris, ethylenediaminetetraacetic acid (EDTA) and vitamin E and found that 100 µg of antioxidant potential is equivalent to 247 µg of EDTA and 58 µg of vitamin. Russo et al. (2003) reported antioxidant potential using DPPH method and reduced hydrogen peroxide-induced cytotoxicity of methanolic extract of whole plant. Administration of 10 mg/kg aqueous bacoside A remarkably enhanced levels of glutathione, Vitamin C, Vitamin A and Vitamin E in brain of rats exposed to cigarette smoke. Bacoside A has also been found to increase superoxide dismutase, catalase, glutathione peroxide and glutathione reductase (Anbarasi et al. 2006). Dey et al. (2019) noticed higher levels of total bacoside content, total phenol content and antioxidant potential in in vitro grown plants treated with polyamines as compared to mother plant. Free radicals produced during the normal metabolic reaction of aerobic cells are extremely reactive and at unsteady state that have tendency to react with food lipids, nucleic acid, sugars and sterols (Lee et al. 2004), which lead to various physiological modifications. Chemical elicitation treatment significantly enhanced free radical scavenging potential of extract.

Ahmad et al. (2013) reported that enhanced phenol content leads to increased antioxidant capacity. Saxena et al. (2016) reported that phenol content (using Folin-Ciocalteu reagent) present in methanolic and ethnolic extract of plants *Eclipta* and *Plumbago* is responsible for antioxidant potential. It was observed that an increase or decrease in total bacosides, phenol content and antioxidant potential was not influenced by growth parameters. All the tested elicitors, when used separately, stimulated yield of total bacosides, total phenols and antioxidant potential, with a significant difference (Fig. 4). Total bacosides and phenol content were much higher with JA-elicited plants while elicitation with SA showed highest increase in antioxidant potential. Antioxidant activity for CUPRAC assay showed highest 231 µg/g GAE (IC 554588), which was enhanced to 1119.8, 1762 and 563 µg/g GAE by the application of elicitors JA, SA and ME respectively (Fig. 4c). Antioxidant activity of accession IC554588 through DPPH assay was noticed 7.15 µg/g GAE, which was increased to 37.6 µg/g, 56.5 µg/g and 14.6 µg/g GAE with the elicitation of JA, SA and ME respectively (Fig. 4d). The present study clearly indicates that elicitation increases accumulation of total

Table 3 Linear correlation between moisture content, antioxidant potential using CUPRAC and DPPH method, bacosides and phenol content of 13 *B. monnieri* accessions

Moisture	CUPRAC	DPPH	Bacosides	Phenol	
Moisture					
Pearson correlation	1	−0.815**	−0.820**	−0.574**	−0.832**
Sig. (2-tailed)		0.000	0.000	0.000	0.000
N	48	48	48	48	48
CUPRAC					
Pearson correlation	−0.815**	1	0.994**	0.811**	0.951**
Sig. (2-tailed)	0.000		0.000	0.000	0.000
N	48	48	48	48	48
DPPH					
Pearson correlation	−0.820**	0.994**	1	0.798**	0.954**
Sig. (2-tailed)	0.000	0.000		0.000	0.000
N	48	48	48	48	48
Bacosides					
Pearson correlation	−0.574**	0.811**	0.798**	1	0.871**
Sig. (2-tailed)	0.000	0.000	0.000		0.000
N	48	48	48	48	48
Phenol					
Pearson correlation	−0.832**	0.951**	0.954**	0.871**	1
Sig. (2-tailed)	0.000	0.000	0.000	0.000	
N	48	48	48	48	48

**Correlation is highly significant at an error of $p < 0.001$ level

bacosides, TPC and antioxidant potential of in vitro cultures of *B. monnieri*.

According to statistical analysis (MLR), antioxidant potential is highly correlated to total bacoside and total phenols as R-sq(adj) is 65.8% and 90.1% respectively. It is also evident that phenol is more correlated to antioxidant potential than bacoside (statistical study). As per regression equation, the coefficient of phenol is higher than that of bacoside, which employs that antioxidant potential, is more sensitive to phenol as compared to bacoside. The sign of coefficients indicate that both the factors are directly correlated to antioxidant potential. The *R*-adj value for the regression equation is 89%, which shows very strong correlation of both the factors with antioxidant potential.

The present study is a comprehensive study of 13 accessions for bacoside, phenol and antioxidant potential, which were collected from different geographical locations. Elite accessions were analyzed on the basis of various parameters such as maximum number of shoots, highest shoot length, total yield of bacosides, TPC and antioxidant potential. Apart from this, JA, SA and ME were used for elicitation which have not been tested for the studied accessions by varying experimental conditions. Elicitation with JA, SA and ME has been performed on all the above tested parameters and noticeable increase was found in amount of bacoside, phenol content and antioxidant potential, which is undoubtedly enhancing their therapeutic potential.

Table 4 *B. monnieri*: eigenvalues for all the five components in 13 accessions

Component	Initial eigen values		
	Total	% of variance	Cumulative %
Moisture	2.816	40.232	40.232
CUPRAC	1.725	24.643	64.875
DPPH	1.261	18.016	82.891
Bacosides	0.628	8.971	91.862
Phenols	0.432	6.172	98.034

Extraction method: principal component analysis

Table 5 Component matrix of five components in 13 accessions

	Component		
	PC1	PC2	PC3
Moisture	0.584	−0.679	0.229
CUPRAC	0.644	0.692	−0.095
DPPH	0.689	0.419	0.431
Bacosides	−0.015	0.431	0.725
Phenol	0.448	0.453	−0.683

Extraction method: principal component analysis

PC1 principal component 1, *PC2* principal component 2, *PC3* principal component 3

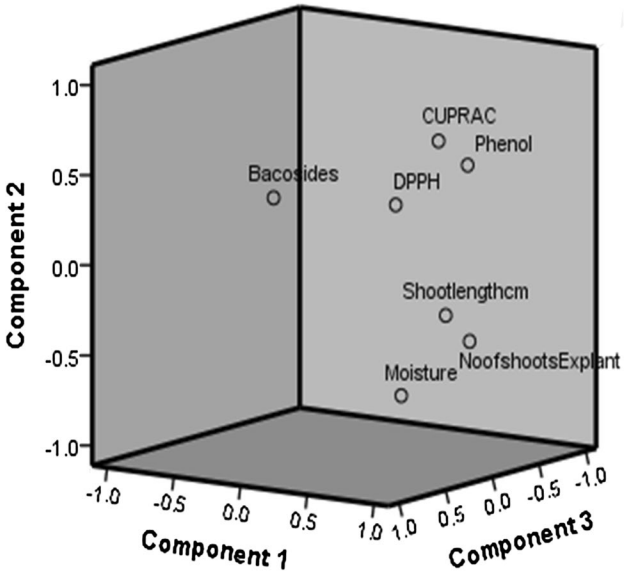


Fig. 5 Three dimensional plot for PCA

Correlation, variation, principal component and cluster analysis

Statistical analysis revealed negative correlation of moisture content with bacoside content, antioxidant potential and TPC (Table 3), which is beneficial as high moisture content can lower the pharmacological effect of plant. Antioxidant by CUPRAC method exhibited strong correlation with DPPH, bacoside and phenol content. Phenols and bacosides are highly correlated with strong positive correlation. Among all components phenol is the major contributor to variation. Elicited samples exhibited increase in bacoside content as well as phenol component.

PCA was carried out to find total variance among all samples. In the present study, five parameters have been taken into account for the same that are moisture content, total bacosides, TPC and antioxidant potential with DPPH and CUPRAC methods. PCA revealed that moisture, CUPRAC, and DPPH together governed 83% of total variability among all five components (Table 4). Therefore,

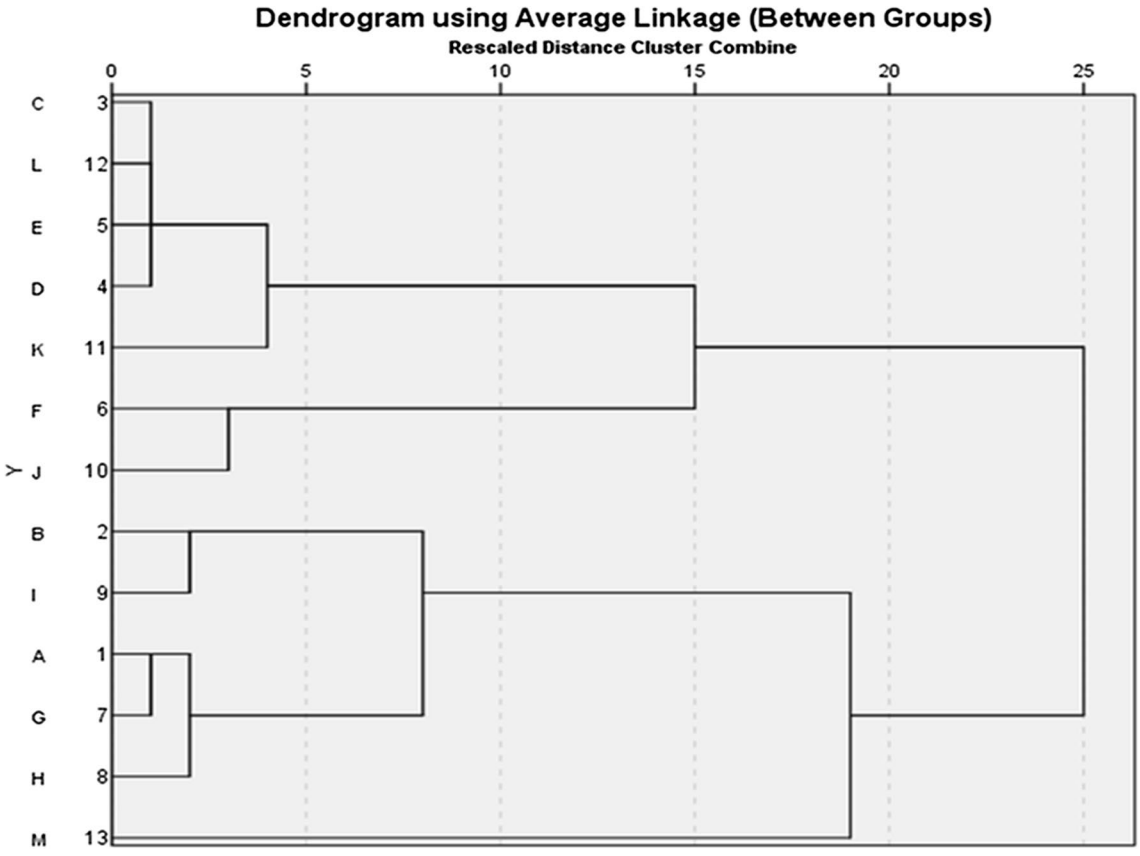


Fig. 6 Clustering of 13 accessions of *B. monnieri* based on components moisture content, antioxidant potential using CUPRAC and DPPH methods, bacosides and phenol content, number of shoots/explants and shoot length. (Refer Table 1 for the corresponding accessions of the shown numbers in the Fig. 4

these three components are considered as Principal Components (PC). Component matrix is shown in Table 5 where PC1 represents moisture while PC2 and PC3 represent CUPRAC and DPPH respectively. In PC1, CUPRAC (0.644) and DPPH (0.689) have high positive loading and are considered as major contributors to variability. In PC2, CUPRAC is major contributor followed by phenol and bacoside. Maximum contribution of bacoside was noticed in PC3. Antioxidant potential as assessed by CUPRAC and DPPH methods are major traits responsible for variability amongst all the tested accessions (Table 5). Three dimensional plot for PCA is shown in Fig. 5. Antioxidant potential was major contributor in cluster grouping, and its influence is distinctly reflected in the primary grouping of 13 *B. monnieri* accessions.

The dendrogram generated, based on judged parameters revealed five distinct clusters containing five accessions in Cluster I, two accessions in each Cluster II and Cluster III, three in Cluster IV and one accession in Cluster V at the linkage distance of 10 (Fig. 6). In general, accessions present in Cluster I, Cluster IV and Cluster V have moderate amount of bacoside content. Accessions of Cluster II showed the highest amount of bacoside while accessions of Cluster III had lowest amount. Accessions of Cluster II (IC 344312 and IC 554588), besides having highest bacoside content also had high TPC and antioxidant potential. IC554585 of Cluster V showed high antioxidant potential (by CUPRAC method) and phenol content with moderate amount of bacoside. Hence, amongst the 13 studied accessions, IC 554588, IC 344312 and IC 554585 are best suited for exploitation for pharmaceutical purposes.

The present study has been conducted with in vitro cultures of *B. monnieri* to enhance the bacoside production, TPC and antioxidant potential, which contributes to its therapeutic value. However, it is an easy growing plant but wild collection from its natural habitat by various pharmaceutical companies has resulted in overexploitation. Use of in vitro cultures and elicitors to enhance pharmaceutically important compounds in culture conditions can protect the natural habitat of this medicinally important plant contributing to its sustainability in the environment.

Conclusion

The study revealed wide variability among thirteen accessions of *B. monnieri*, collected from different states of India. Variability was defined with respect to total bacoside content, TPC and antioxidant potential. Highly significant and strong correlation was found between total bacosides, total phenols and antioxidant activity, which were the major contributors in cluster grouping. IC 554588, IC 344312 and IC 554585 have been identified elite accessions with higher bacoside production, TPC and antioxidant potential as compared to others. Use of elicitors enhanced the total bacosides and TPC, which are responsible for higher antioxidant potential. Hence, elicitor-treated and superior accessions can be cultivated by farmers to increase the raw material for drug manufacturing.

Acknowledgements We are grateful to Vice Chancellor of Delhi Technological University, Delhi, India for constant support and encouragement in conducting this study. We are also thankful to Dr. Ram Singh, Assistant Professor (Applied Chemistry) and his advisee Dr. Atya and Dr. Gitika (Applied Chemistry), Delhi Technological University for their help in performing HPLC.

Authors contribution statement Experimental strategy belongs to NB, NS and NJ. NJ and RB performed experiment. RB did statistical analysis.

Funding This research did not receive any grant from funding agencies in the public, commercial, or not-for-profit sectors.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Ahmad S, Garg M, Tamboli ET, Abdin MZ, Ansari SH (2013) In vitro production of alkaloids: factors, approaches, challenges, and prospects. Pharmacogn Rev 7:27–33. https://doi.org/10.4103/0973-7847.112837

Alamgir ANM, Rahman A, Rahman M (2014) Secondary metabolites and antioxidant activity of the crude leaf extract of *Bacopa monniera* (L.) Pennel. and *Coccinia grandis* (L.). J Pharmacog Phytochem 3:226–230

Ali M, Yu KW, Hahn EJ, Paek KY (2006) Methyl jasmonate and salicylic acid elicitation induce ginsenosides accumulation, enzymatic and non-enzymatic antioxidant in suspension culture *Panax ginseng* roots in bioreactors. Plant Cell Rep 25:613–620. https://doi.org/10.1007/s00299-005-0065-6

Anbarasi K, Vani G, Balakrishna K, Desai CS (2005) Creatine kinase isoenzyme patterns upon chronic exposure to cigarette smoke: protective effect of Bacoside A. Vascul Pharmacol 42:57–61. https://doi.org/10.1016/j.vph.2005.01.003

Anbarasi K, Vani G, Balakrishna K, Devi CS (2006) Effect of bacoside A on brain antioxidant status in cigarette smoke exposed rats. Life Sci 78:1378–1384

Apak R, Guclu K, Ozyurek M, Karademir SE (2004) Novel total antioxidant capacity index for dietary polyphenols and Vitamins A and E, using their cupric ion reducing capability in the presence of neocuproine: CUPRAC method. J Agri Food Chem 52:7970–7981. https://doi.org/10.1021/jf048741x

Atoui AK, Mansouri A, Boskou G, Kefalas P (2005) Tea and herbal infusions: their antioxidant activity and phenolic profile. Food Chem 89:27–36. https://doi.org/10.1016/j.foodchem.2004.01.075

Basu N, Rastogi RP, Dhar ML (1967) Chemical examination of *Bacopa monniera* Wettst: part III, bacoside B. Indian J Chem 5:84–86

Brand-Williams W, Cuvelier ME, Berset C (1995) Use of free radical method to evaluate antioxidant activity. Lebensmittel Wissenschaft Technologie 28:25–30

Chen CH, Pearson AM, Gray JI (1992) Effects of synthetic antioxidants (BHA, BHT, and PG) on the mutagenicity of IQ-like compounds. Food Chem 43:177–183

Chen CH, Chan CH, Chu YT, Ho HY, Chen PY, Lee TH, Lee CK (2009) Antioxidant activity of some plant extracts towards xanthine oxidase, lipoxygenase, and tyrosinase. Molecules 14:2947–2958. https://doi.org/10.3390/molecules14082947

Dey A, Hazra AK, Nongdam P, Nandy S, Tikendra L, Mukherjee A, Banerjee S, Mukherjee S, Pandey DK (2019) Enhanced bacoside content in polyamine treated in vitro raised *B. monnieri* (L.) Wettest. South Afr J Bot 123:259–269

Farmer EE, Almeras E, Krishnamurthy V (2003) Jasmonates and related oxylipins in plant responses to pathogenesis and herbivory. Curr Opin Plant Biol 6:372–378

Gorelick J, Bernstein N (2014) Elicitation: an underutilized tool for the development of medicinal plants as a source for therapeutic secondary metabolites. Adv Agron 124:201–230. https://doi.org/10.4103/0973-1296.185726

Gottlieb OR, Borin MR (2000) Medicinal products: regulation of biosynthesis in space and time. Mem Inst Oswaldo Cruz 95:115–120

Hazra S, Bhattacharyya D, Chattopadhyay S (2017) Methyl jasmonate regulates podophyllotoxin accumulation in podophyllum hexandrum by altering the ros-responsive podophyllotoxin pathway gene expression additionally through the down regulation of few interfering miRNAs. Front Plant Sci 8:1–9. https://doi.org/10.3389/fpls.2017.00164

Hussain MS, Fareed S, Ansari S, Rahman MA, Ahmad IZ, Saeed M (2012) Current approaches toward the production of secondary plant metabolites. J Pharm Bioallied Sci 4:10–20. https://doi.org/10.1007/s10499-014-9870-7

Jain P, Sharma HP, Basri F, Priya K, Singh P (2017) Phytochemical analysis of *Bacopa monnieri* (L.) Wettst. and their antifungal activities. Indian J Tradit Knowl 16:310–318

Jauhari N, Bharadvaja N, Sharma N (2016) One step to conserve medicinally important plant *Bacopa monnieri* through rapid and cost-effective in vitro propagation. Progress Agric 16:8–15. https://doi.org/10.5958/0976-4615.2016.00002.8

Kapoor LD (1990) CRC handbook of ayurvedic medicinal plants. CRC Press, Boca Raton

Karuppusamy S (2009) A review on trends in production of secondary metabolites from higher plants by in vitro tissue, organ and cell cultures. J Med Plants Res 3:1222–1239

Kumar N, Abichandani LG, Thawani V, Gharpure KJ, Naidu MU, Venkat RG (2016) Efficacy of standardized extract of *Bacopa monnieri* (Bacognize®) on cognitive functions of medical students: a six-week, randomized placebo-controlled trial. Evid Based Complement Alternat Med. https://doi.org/10.1155/2016/4103423

Kundu K, Roy A, Saxena G, Kumar L, Bharadvaja N (2016) Effect of different carbon sources and elicitors on shoot multiplication in accessions of *Centella asiatica*. Med Aromat Plants 5:251–255. https://doi.org/10.4172/2167-0412.1000251

Largia J, Pothiraj G, Shilpa J, Ramesh M (2015) Methyl jasmonate and salicylic acid synergism enhance bacoside A content in shoot cultures of *Bacopa monnieri* (L.). Plant Cell Tiss Organ Cul 122:9–20. https://doi.org/10.1007/s11240-015-0745-z

Lee J, Koo N, Min DB (2004) Reactive oxygen species, aging, and antioxidative Nutraceuticals. Compr Rev Food Sci Food Saf 3:21–33. https://doi.org/10.1111/j.1541-4337.2004.tb00058.x

Leonard J, Seth B, Sahu BB, Singh VR, Patra N (2018) Statistical optimization for enhanced bacoside A production in plant cell cultures of B. monneiri. Plant Cell Tiss Organ Cult 133:203–214

Madhu K, Prakash T, Maya S (2019) Bacoside-A inhibits inflammatory cytokines and chemokine in the experimental autoimmune encephalomyelitis. Biomed Pharmacoth 109:1339–1345

Mangas S, Moyano E, Hernández-Vázquez L, Bonfill M (2009) *Centella asiatica* (L) Urban: an updated approach. Plant secondary terpenoids. Research Signpost, Trivandrum, pp 55–74

Mangathayaru K (2013) Pharmacognosy: an Indian perspective, 1st edn. Pearson

Meena H, Pandey HK, Pandey P, Arya MC, Ahmed Z (2012) Evaluation of antioxidant activity of two important memory enhancing medicinal plants *Bacopa monnieri* and *Centella asiatica*. Indian J Pharmacol 44:114–117. https://doi.org/10.4103/0253-7613.91880

Mendoza D, Cuaspud O, Arias JP, Ruiz O, Arias M (2018) Effect of salicylic acid and methyl jasmonate in the production of phenolic compounds in plant cell suspension cultures of *Thevetia peruviana*. Biotechnol Rep (Amst) 19:e00273. https://doi.org/10.1016/j.btre.2018.e00273

Mishra A, Mishra AK, Tiwari OP, Jha S (2013) HPLC analysis and standardization of Brahmi vati—an ayurvedic poly-herbal formulation. J Young Pharm 5:77–82. https://doi.org/10.1016/j.jyp.2013.09.001

Mukherjee S, Dugad S, Bhandari R, Pawar N, Jagtap S, Pawar PK, Kulkarni O (2011) Evaluation of comparative free-radical quenching potential of Brahmi (*Bacopa monnieri*) and Mandookparni (*Centella asiatica*). Ayu 32:258–264. https://doi.org/10.4103/0974-8520.92549

Mundkinajeddu D, Sangli GK, Chandrasekharappa AP, Agarwal A (2005) Quantitative determination of the major saponin mixture Bacoside A in *Bacopa monnieri* by HPLC. Phytochem Anal 16:24–29. https://doi.org/10.1002/pca.805

Murashige T, Skoog F (1962) A revised medium for rapid growth and bio assays with tobacco tissue cultures. Physiol Plant 15:473–497. https://doi.org/10.1111/j.1399-3054.1962.tb08052.x

Murthy PBS, Raju VR, Ramakrisana T, Chakravarthy MS, Kumar KV, Kannababu S, Subbaraju GV (2006) Estimation of twelve bacopa saponins in Bacopa monnieri extracts and formulations by high-performance liquid chromatography. Chem Pharm Bull 54:907–911

Namdeo AG (2007) Plant cell elicitation for production of secondary metabolites: a review. Phcog Rev 1:69–79

National Medicinal Plant Board (2019) Government of India, New Delhi, India. Accessed May 2019

Pal R, Dwivedi AK, Singh S, Kulshrestha DK (1998) High performance liquid chromatographic (HPLC) quantification of bacosides in Bacopa monnieri extracts. Indian J Pharm Sci 60:328–329

Pawar R, Gopalakrishnan C, Bhutani KK (2001) Dammarane triterpene saponin from *Bacopa monniera* as the superoxide inhibitor in polymorphonuclear cells. Planta Med 67:752–754. https://doi.org/10.1055/s-2001-18351

Rahnamaie-Tajadod R, Loke K, Goh H, Noor NM (2017) Differential gene expression analysis in polygonum minus leaf upon 24 h of methyl jasmonate elicitation. Front Plant Sci 8:1–14. https://doi.org/10.3389/fpls.2017.00109

Roodernrys S, Booth D, Bulzomi S, Phipps A et al (2002) Chronic effects of Brahmi (*Bacopa monnieri)* on human memory. Neuropsychopharmacology 27:279–281. https://doi.org/10.1016/S0893-133X(01)00419-5

Russo A, Izzo AA, Borrelli F, Renis M, Vanella A (2003) Free radical scavenging capacity and protective effect of Bacopa monniera L. on DNA damage. Phytother Res 17:870–875. https://doi.org/10.1002/ptr.1061

Saeed S, Ali H, Khan T, Kayani W, Khan MA (2017) Impacts of methyl jasmonate and phenyl acetic acid on biomass accumulation and antioxidant potential in adventitious roots of *Ajuga bracteosa* Wall ex Benth., a high valued endangered medicinal plant.

Physiol Mol Biol Plants 23:229–237. https://doi.org/10.1007/s12298-016-0406-7

Sairam K, Dorababu M, Goel RK, Bhattacharya SK (2002) Antidepressant activity of standardized extract of *Bacopa monniera* in experimental models of depression in rats. Phytomedicine 9:207–211. https://doi.org/10.1078/0944-7113-00116

Saxen A, Gautam S, Arya KR, Singh RK (2016) Comparative study of phytochemicals, antioxidant potential and activity of enzymatic antioxidants of *Eclipta alba* and *Plumbago zeylanica* by in vitro assays. Free Rad Antiox 6:139–144. https://doi.org/10.5530/fra.2016.2.2

Shabani L, Ehsanpour A, Asghari G, Emami J (2009) Glycyrrhizin production by in vitro cultured *Glycyrrhiza glabra* elicited by methyl Jasmonate and salicylic acid. Russ J Plant Physol 56:621–626

Shahid M, Subhan F, Ahmad N, Ullah I (2017) A bacosides containing *Bacopa monnieri* extract alleviates allodynia and hyperalgesia in the chronic constriction injury model of neuropathic pain in rats. BMC Complement Altern Med 17:293–305. https://doi.org/10.1186/s12906-017-1807-z

Sharma N, Satsangi R, Pandey R, Devi SV (2007) In vitro clonal propagation and medium term conservation of Brahmi [*Bacopa monnieri* (L.) Wettst]. J Plant Biochem Biotech 16:139–143. https://doi.org/10.1007/978-1-4939-3332-7_11

Sharma N, Satsangi R, Pandey R, Singh R, Kaushik N, Tyagi RK (2012) In vitro conservation of *Bacopa monnieri* (L.) using mineral oil. Plant Cell Tiss Organ Cult 111:291–301. https://doi.org/10.1007/s11240-012-0194-x

Sharma P, Yadav S, Srivastava A, Srivastava N (2013) Methyl jasmonate mediates upregulation of bacoside A production in shoot cultures of *Bacopa monnieri*. Biotech Lett 35:1121–1125. https://doi.org/10.1007/s10529-013-1178-6

Sharma M, Ahuja A, Gupta R, Mallubhotla S (2015) Enhanced bacoside production in shoot cultures of *Bacopa monnieri* under the influence of abiotic elicitors. Nat Prod Res 29:745–749. https://doi.org/10.1080/14786419.2014.986657

Singleton VL, Orthofer R, Lamuela-Raventos RM (1999) Analysis of total phenols and other oxidation substrates and antioxidants by means of Folin-Ciocalteu reagent. Methods Enzymol 299:265–275. https://doi.org/10.1016/S0076-6879(99)99017-1

Thaipong K, Boonprakob U, Crosby K, Cisneros-Zevallos L, Byrne DH (2006) Comparison of ABTS, DPPH, FRAP, and ORAC assays for estimating antioxidant activity from guava fruit extracts. J Food Compos Anal 19:669–675. https://doi.org/10.1016/j.jfca.2006.01.003

Tripathi YB, Chaurasia S, Tripathi E, Upadhyay A, Dubey GP (1996) Bacopa monniera Linn. as an antioxidant: mechanism of action. Indian J Exp Biol 34:523–526

Verma P, Mathur AK, Jain SP, Mathur A (2012) In vitro conservation of twenty-three overexploited medicinal plants belonging to the Indian subcontinent. Sci World J 15:85. https://doi.org/10.1100/2012/929650 **(Article ID 929650)**

Walton NJ, Brown DE (1999) Chemicals from plants: perspectives on plant secondary product. Imperial College Press, London

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

Materials Research Express

ACCEPTED MANUSCRIPT

Development of Polypropylene/Banana Peel (treated and untreated) composites with and without compatibilizer and their studies

To cite this article before publication: Priyanka Luthra *et al* 2019 *Mater. Res. Express* in press https://doi.org/10.1088/2053-1591/ab2eea

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 08/07/2019 at 05:40

Development of Polypropylene/Banana Peel (Treated and Untreated) Composites with and without Compatibilizer and their Studies

Priyanka Luthra[1,2], Ram Singh[1]* and Gurpreet Singh Kapur[2]*

[1]*Department of Applied Chemistry, Delhi Technological University, Delhi-110042, India*
[2]*Indian Oil Corporation Ltd., Research & Development Centre, Sector 13, Faridabad-121007, Haryana, India*
ramsingh@dtu.ac.in and kapurgs@indianoil.in

Abstract

The use of agricultural waste in the preparation of polymer composites is a step towards the replacement of synthetic fibers with the natural fibers. In the present study, banana peel (BP) has been explored as a source of natural fiber for the preparation of composites. The effect of treated (sodium hydroxide, potassium permanganate, silane and Lacasse enzyme treatment) and untreated BP fiber on polypropylene (PP) matrix have been studied using the fiber loading of 0, 10, 20, 30, and 40 wt%. The effect of compatibilizer (maleic anhydride grafted polypropylene [PP-g-MA], 1, 2 and 4 wt%) on the properties have also been studied with selected natural filler loading. The prepared composites have been evaluated for morphology, mechanical, thermal, rheological and water absorption properties. The results have showed an improvement in different properties with increase in the content of banana peel for PP/BP composites.

Keywords: Banana peel, fiber, composites, polypropylene, morphology, compatibilizers

Accepted Manuscript

1. Introduction

The natural fiber composite is a step towards moving from the traditional composites to green composites which showed better compatibility with the environment [1-5]. Natural fiber reinforced polymer composites showed specific properties and specialized applications in varied industries like automotive, construction, transportation and so on [6-8]. Due to the better availability, low cost, flexibility and environmentally friendly properties, the natural fibers make themselves superior to synthetic fibers [9-10]. Several studies have been reported towards the use of natural fibers as fillers like *calatropsis gigantea* powder, tea waste, *azadirachta indica* seed powder, etc and their applications in composite formation [11-14]. The *calotropis gigantea* stem fiber was used with epoxy to develop hybrid composite [12]. In another study, untreated tea mill waste fibers were used with matrix high density polyethylene and polypropylene [13]. The untreated prawn shell powder with epoxy gave composite material that showed high strength, tough, and good hardness to withstand loads and water resistance [14]. The utilization of natural fiber mainly from agricultural waste has been given more attention in recent time [15-17].

Banana is one of the important fruits consumed worldwide, but its non-edible peel (Figure S1) is a solid waste and pose a threat to the environment. India is the largest producer of banana in the world with 29.7 million tonnes [18], and hence its solid waste requires special attention. About 12% of the total weight of the banana accounts the weight of its peel. Banana peel as trash is dumped in the municipal landfill.

Banana peel is chemically composed of pectin, cellulose, hemicellulose and small amounts of lignin [19]. There have been reports of using banana peel as reinforcing agents to make composites. Pereria et al prepared banana peels composites using washed and unwashed peels with high-density polyethylene (HDPE) as matrix [20]. Naidu et al. studied banana peel reinforced epoxy composites [21]. Akbay et al studied the use of banana peel in ethylene propylene diene rubber (EPDM) as an alternate to carbon black in tyres [22]. To the best of our knowledge, no work has been cited on the loading of banana peel in PP matrix and exceeding the limit of 30wt%. The study on the treatment of banana peel and their applications for the preparation of composites have also been not conducted. In present work, the effect of banana peel fibrous powder (BP) content, treated and untreated both, on the morphological, mechanical, visco-elastic behavior and thermal properties of PP is investigated. These kinds of composites will find its application in various sectors like automotive, construction etc due to its advantages

Accepted Manuscript

of high strength, resistance to corrosion, good impact strength, strength related to weight, dimensional stability, flexibility in design. These composites are a suitable alternate to synthetic fiber-based composites.

2. Materials and methods

2.1 Materials

Isotactic polypropylene (PP) powder manufactured by IOCL, with a melt flow index of 11g/10 min at 230^{o}C under load of 2.16 kg was used as the matrix. PP powder was used after being stabilized by primary and secondary antioxidants. The maleic anhydride grafted polypropylene known as polybond 3200 (PP-g-MA) having MFI 115g/10 min at 190^{o}C under load of 2.16 kg was used as compatibilizer and purchased from M/s Addivant, USA. The characteristic properties of PP and PP-g-MA as per supplier datasheet are given in Table S1. Banana Peel fibrous powder (BP) was obtained from M/s Herbs & Crops Overseas, India. It was sieved through a mesh of 106 microns for the studies.

2.2 Preparation of PP/BP composites without compatibilizer

Composites were prepared by melt blending using Mini Lab II Haake Rheomex CTW5, co-rotating twin screw extruder. Prior to processing, BP was oven dried at 80 °C for 6 h. The melt from the extruder was filled into the barrel and injection molded at an injection pressure of 500 bar and mold temperature 60°C. The temperature of the extruder and the barrel of injection molding were kept at 200°C. Test samples were prepared using Haake Minijet II injection molding machine. Screw rpm was kept around 150 rpm and compounding and molding temperature kept were 200°C. The samples have been designated as PPxB, where “PP” stands for polypropylene, “B” stands for BP fiber and “x” is the amount of BP in weight % (Table 1).

2.3 Preparation of PP/BP composites with compatibilizer

Similar processing conditions (section 2.2) are used for the preparation of PP/PP-g-MA/BP composites using PP-g-MA as compatibilizer. In all these samples, the amount of untreated BP was kept constant i.e. 20 wt% and compatibilizer amount was varied in the ratio of 1, 2 and 4 wt%. After evaluation of percentage of compatibilizer an optimized concentration of the same i.e. 1 wt% was added and evaluated with 30 and 40wt% of BP fiber in PP matrix. The

Accepted Manuscript

compatibilizer loaded samples are designated as PPxByC where "y" stands for the percentage of compatibilizer and "C" is for the compatibilizer (Table 1).

2.4 Treatments of BP fibers

Chemical treatments

Alkaline treatment: BP fibers were soaked in aqueous sodium hydroxide solution (5%, w/v) (fiber: solution, 1:20, g/mL) at 23°C temperature for 24 h under continuous stirring. After treatment, fibers were thoroughly washed with distilled water to remove any traces of alkali on the fiber surface. The treated fibers (BA) were further dried and used for composite preparation.

Potassium permanganate treatment: Alkali treated BP fibers were soaked in $KMnO_4$/acetone solution (0.5%, w/v) (fiber: solution, 1:15, g/mL) for 10 minutes with continuous stirring. After treatment, fibers were thoroughly washed with distilled water. The treated BP fibers (BK) were dried and used for composite preparation.

Silane treatment: The BP fiber was treated with 3-aminopropyltriethoxy silane solution in isopropanol/water (2:3 v/v) (1%, w/v) (fiber:solution,) for 2 h. The reaction mixture was filtered, the treated BP powder was dried and labeled as BS.

Enzyme treatment: The enzyme *Laccase* (30 u/g) was used to treat the BP fiber for 2, 4 and 6 h under continuous stirring at 23°C temperature and 5.5 pH [23]. A fiber to solution ratio of 1:30 (g/mL) was maintained for the experiment. After appropriate time, the fibers were washed with distilled water until a pH close to 7.0 of filtrate was reached for avoiding any acidic effect on them. The treated BP fibers were dried and labeled as BE-z where z represents the treatment time from 2 to 6 hours. The obtained BP fibers were used for composite preparation similar to sections 2.2 and 2.3.

The amount of treated BP was kept constant i.e. 30 wt% and 40 wt% and compatibilizer amount was kept as 1wt%. The detail of the sample designation of all the prepared composites is given in Table 1.

3. Results and discussion

3.1 Analysis and characterization of banana peel (BP) fibers

Compositional analysis of BP fiber was carried out by following the Laboratory Analytical Procedure (LAP) of the National Renewable Energy Laboratory (NREL). The results

Accepted Manuscript

showed the presence of cellulose (34.6%), hemicellulose (18.2%), lignin (16.8%), acid insoluble ash (1.4%) along with water extractives (24.5%) and ethanol extractives (4.31%) in untreated BP fiber [19]. The treated BP fibers showed a decrease in each component [24]. The elemental composition (C, H, N, & S) and bulk density are given in Table S2. The bulk density of the fibers decreases after the treatment infer the decrease in lignin and hemicellulose content. The structural changes in BP fibers before and after treatments were further characterized using FTIR, TGA, SEM, and XRD.

3.1.1 FTIR analysis of BP fibers

The FTIR spectra of untreated and treated BP fibers are shown in Figure 1(a,b). The absorptions for untreated BP at 3381, 2920, 1733, 1600 and 1247 cm^{-1} are respectively ascribed to the OH stretching, C–H stretching, C=O and C–O stretching of cellulose, hemicelluloses and lignin groups [25,26].

Table 1. PP and PP/BP composites along with their sample designations

Sample[#]	PP	BP	PP-g-MA	**Sample**[#]	PP	BP	PP-g-MA
PP	100	0	0	PP30BK	69	30	1
PP10B	90	10	0	PP30BS	69	30	1
PP20B	80	20	0	PP30BE-2	69	30	1
PP30B	70	30	0	PP30BE-4	69	30	1
PP40B	60	40	0	PP30BE-6	69	30	1
PP20B1C	79	20	1	PP40BA	59	40	1
PP20B2C	78	20	2	PP40BK	59	40	1
PP20B4C	76	20	4	PP40BS	59	40	1
PP30B1C	69	30	1	PP40BE-2	59	40	1
PP40B1C	59	40	1	PP40BE-4	59	40	1
PP30BA	69	30	1	PP40BE-6	59	40	1

[#]*PP:polypropylene, B: BP fiber, C: compatibilizer, BA: alkali treated fiber, BK: potassium permanganate treated fiber, BS: silane treated fiber, BE: Lacasse enzyme treated fiber, 2, 4 and 6 stands for time duration of enzyme treatment*

After treatment, the absorption at 1733 cm^{-1} for C=O bond of the carboxylic group of hemicelluloses and pectin was observed to be decreasing, which confirmed the decrease in the

Accepted Manuscript

concentration of the constituents due to the treatment of fiber. The treatment either removed or lower the amount of lignin as inferred from the absence of the peak at 1247 cm^{-1} in treated BPs that corresponds to C–O stretching vibration of acetyl groups of lignin [23]. A distinct peak was observed at 1413 cm^{-1} in treated fibers which were a shoulder in untreated fibers, showed the presence of cellulose. An intensity of the small peak at 891 cm^{-1} has increased in the surface modified fibers. The band at this region corresponds to β-1,4-glycosidic bond characteristic absorption between the monosaccharide of cellulose [27].

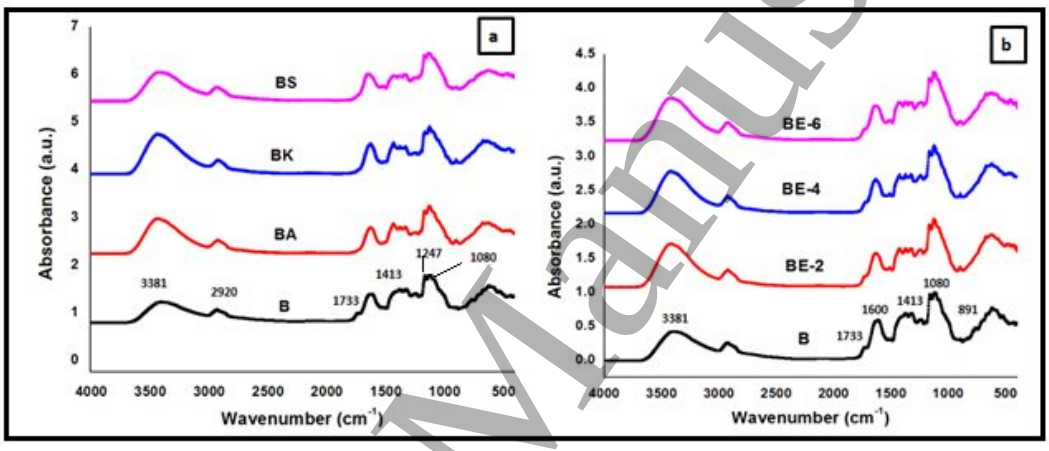


Figure 1. (a) BP-untreated and chemical treated, (b) BP-untreated and enzyme treated at different time interval

3.1.2 Thermogravimetric analysis (TGA) of BP fiber

In TGA, the four-step degradation is observed in the untreated fiber which changes to three-step degradation after treatment (Figure 2(a-d)). This is due to the removal of cementing materials i.e. lignin as the peak at high-temperature region around 430-680°C which was observed in the untreated fibers diminishes. A ~5.3% weight loss up to 100°C was observed in untreated fiber and this attributed to the water loss as the fibers are hydrophilic in nature. Hemicellulose decomposes mainly between 150 and 350°C, cellulose decomposes between 275 and 350°C and degradation of lignin in the temperature range of 350-750°C [28]. The onset of degradation and peak degradation temperature after treatment of fibers increased to ~29-43°C and ~40-49°C respectively which confirms the increase in thermal stability of the fiber as high energy is needed to break the bonds between cellulose and the treating agent/coupling agent

Accepted Manuscript

which is in line with the literature findings [28,29]. After the degradation of fiber due to weight loss of cellulose, hemicelluloses and lignin, residual char is left [29]. The TGA data has been presented in Table S3.

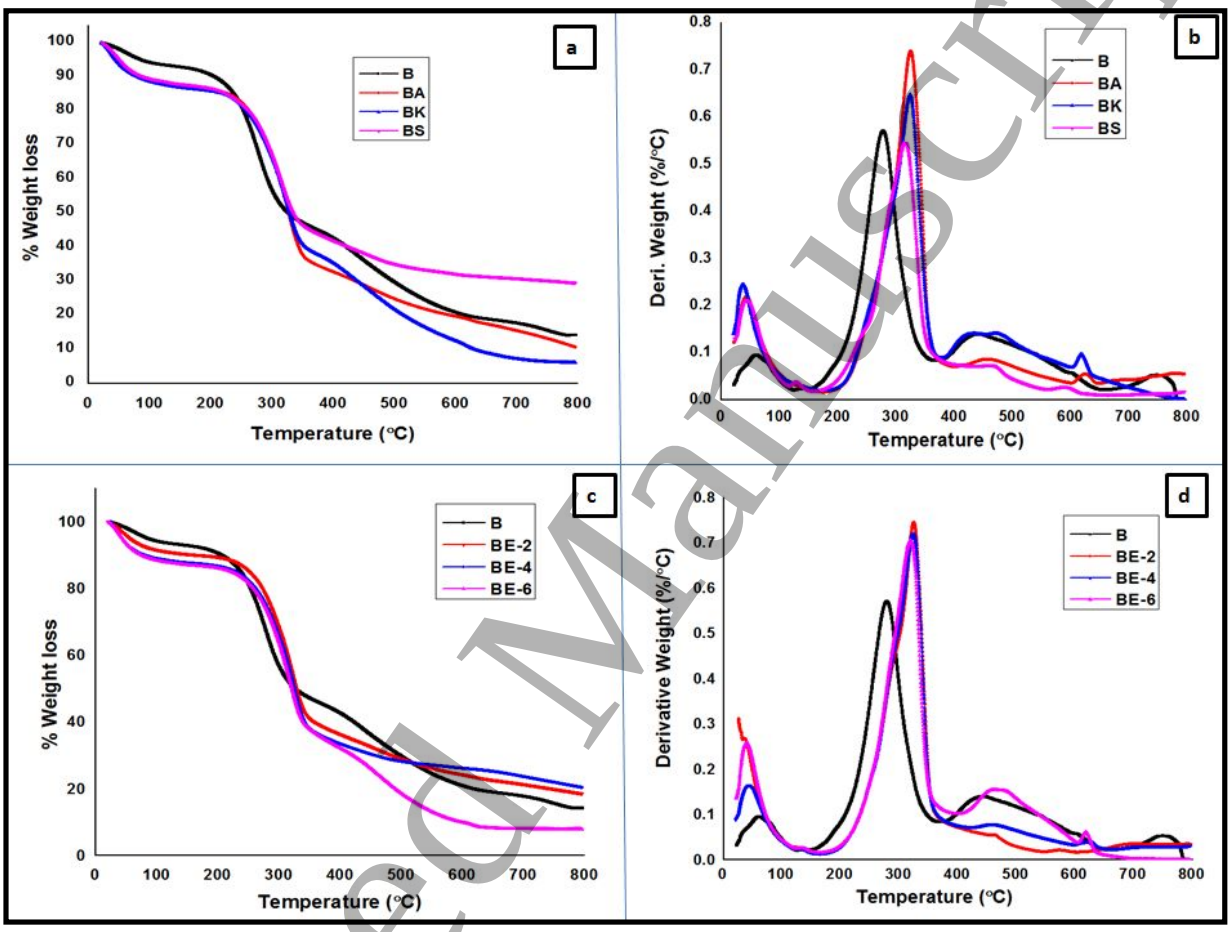


Figure 2. TGA curves for (a) BP & chemically treated BP fibers (c) BP & enzyme treated BP fibers; DTG curves for (b) BP & chemically treated BP fibers (d) BP & enzyme treated BP fibers.

3.1.3 Morphological study with SEM of BP fibers

Surface morphology investigation of untreated BP fiber showed the presence of bundles of fibrils that are bound together by hemicelluloses and lignin (Figure S2(a,b)). This also indicated the coverage of the surface of BP fibers with a layer of substances like oils, waxes, and extractives which agrees with the literature [30]. The shape for the fibers remained cylindrical after treatment with fibrillation observed as seen in the images (Fig. S2(c-h) & Fig. S3(a-f)). In

Accepted Manuscript

comparison to untreated fibers, the treated fibers differ in terms of surface smoothness. The treatment of the fibers made the surface of the fiber rough which could further promote the interlocking adhesion between fiber and the matrix in the preparation of the composites. In the micrographs, pores have been observed in treated fibers which indicated the removal of lignin and hemicelluloses. As observed from SEM micrographs, numbers of pores observed more in case of enzyme treatment carried out for 2 hours duration (Figure S3(a,b)).

3.1.4 X-ray Diffraction (XRD)-Crystallinity Index (CI)

Figures 3(a,b) shows the X-ray diffraction patterns of untreated and various treated BP fibers. As can be observed through the diffractogram, the treatment has led to an increase in the fiber crystallinity index. The major and intense peaks have been observed at values $2\theta = 16^{o}$ & 22^{o}, which represents the cellulose crystallographic planes I_{101} and I_{002} respectively which is same for all the fibers untreated and treated only the intensity varies. The calculations of I_{101}, I_{002} and CI% for each sample were done and summarized in Table 2 as per literature [31].

The PALT fiber peak intensity is the sharpest in comparison to the untreated fiber among all the treatments which confirm that the removal of the amorphous portion thus increasing the crystalline portion of the fiber. In the case of enzyme treatment fiber, with the increase in the treatment time the intensity of the crystalline portion in the peak decreases. This is attributed to crystalline areas disruption due to swelling of the crystalline region of cellulose with an increase in the treatment time and therefore, it reacts with the fiber leading to decreased CI value.

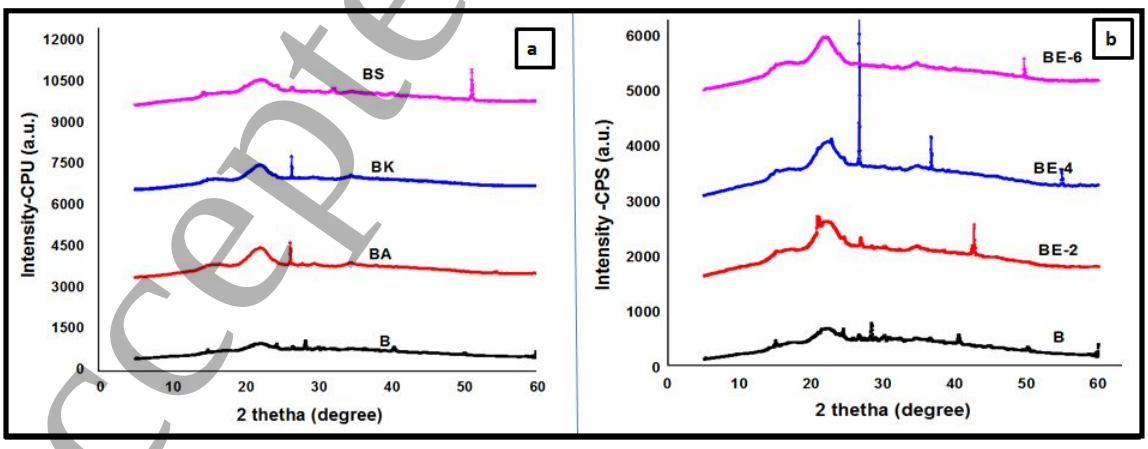


Figure 3. XRD plots for (a) chemical (alkali, permanganate and silane) treated and (b) enzyme treated BP fibers

Accepted Manuscript

Table 2. Diffraction peaks and % crystallinity index of untreated and treated BP fibers

Sample Detail	I_{002}	I_{101}	% Crystallinity Index
B	426 (16.52)	678 (22.12)	49
BA	591 (15.28)	1200 (21.94)	76
BK	568 (16.01)	1068 (22.17)	74
BS	577 (16.3)	1062 (22.15)	65
BE-2	602 (16.4)	1120 (22.05)	73
BE-4	583 (16.2)	1097 (22.20)	60
BE-6	615 (16.4)	1072 (21.4)	59

3.2 Analysis and characterization of PP/BP composites

3.2.1 Morphological study with SEM

SEM micrographs of PP/BP (untreated) composites are shown in Figures S4 (a-g). The presence of holes and a gap between untreated BP and the matrix was observed due to weak adhesion that resulted in the fiber pull out. The fiber pulling out confirmed the poor dispersion of natural fiber into the matrix and weak interfacial bonding between the natural filler and the PP matrix. Agglomeration of natural fibers has been observed in higher loading of fiber indicating the poor dispersion of the filler in the PP matrix. With the incorporation of the compatibilizer, the dispersion of natural fiber was improved which has led to decrease in fiber pull out as shown by SEM (Figures S4(e-g)). These observations agree with the reported studies for natural fiber filled polymer composites [32,33].

The treatment of BP fiber has led to a better dispersion of fiber in PP matrix which thereby confirms the improvement in interfacial adhesion between fiber and the matrix. The SEM micrographs of PP/BP (chemical treated) composites (Figure S5) and PP/BP (enzyme treated) composites (Figure S6) showed no agglomeration of fibers in the composites. With the increase in enzyme treatment time for fiber, the presence of a greater number of voids are observed (Figures S6 (e,f)). This could be due to the degradation of the fiber quality by making the fiber softer. This is in agreement with a previous study [34].

3.2.2 Mechanical properties of PP/BP fibers composites

Accepted Manuscript

Table 3 and Figures S7 and S8 present the tensile strength, tensile modulus, % elongation, flexural strength, flexural modulus and impact strength of PP and PP/BP composites. The incorporation of BP untreated fiber in PP resulted in an enhancement in the tensile modulus with and without PP-g-MA. About 39% and 70% improvement in tensile modulus have been achieved in samples PP30B and PP40B respectively in comparison to PP. The reinforcement action takes place by the transfer of the forces by the matrix to the fibers, so that the fibers can withstand the forces, hence, higher fiber loading can withstand higher forces than lower fiber content [35]. A decrease in tensile strength has been observed with the increase in fiber loading from 10 to 40 wt% in comparison to PP which could be due to agglomeration of fibers [36]. However, better results have been obtained in the case of compatibilized samples (PP30B1C and PP40B1C). With the addition of compatibilizer, an increase in tensile modulus and tensile strength has been observed for a given composition. A hike in modulus affirmed the reinforcing effect of BP fiber which is enhanced further with the addition of compatibilizer which could be due to the better dispersion of filler and enhanced interaction as also supported by the SEM images. Tensile modulus improved by 35% and 13% by incorporation of compatibilizer (PP30BC and PP40BC) in comparison to PP/BP (untreated) composites (PP30B and PP40B). The addition of compatibilizer also further led to the enhancement of tensile strength of PP30B1C (~29%) and PP40B1C (~48%) in comparison to PP30B and PP40B [37].

An irregular change in the flexural strength was obtained in untreated BP fibers-based composite with and without compatibilizer (Table 3, column 5). However, after treatment of fiber, a maximum increase of 35% flexural strength was obtained (PP40BS). In the case of flexural modulus, the composites PP30B and PP40B showed ~16 and 38% increase respectively. The use of compatibilizer further enhances the flexural modulus. The composites with treated BP fibers showed better results in comparison to their untreated counterpart.

The pretreatment of fiber (chemical and enzyme treatment) has led to change in the surface being rougher [38] which is also confirmed by SEM. The increase in mechanical strength and modulus of composite thereby confirms an increase in the concentration of cellulose by removal of lignin and hemicelluloses from the BP fiber [39] which allows better interfacial adhesion between PP and BP fiber. Silane which has been used as a treating agent for the fibers lessens cellulose hydroxyl groups in the fiber which gets hydrolyzed to silanol that leads to condensing with another neighboring silanol group. This further leads to the formation of

Accepted Manuscript

covalent bonding between matrix and the fiber [40]. An increase in enzyme treatment time from 2, 4 and 6 hours of BP fiber has led to drop in the tensile and flexural properties which could be due to the decrease in strength of the fiber due to a long duration of treatment [41]. As expected, % elongation decreased upon incorporation of BP fiber in PP due to the brittle behavior of the PP/BP composites but it increased slightly further on adding PP-g-MA.

Notched izod impact strength

The izod notched impact strength of PP/BP composites increased ~ 52% (PP40B) on increasing BP fiber content which confirms the reinforcement characteristic of the natural fiber added (Table 3, Column 7). A decrease in izod impact properties was observed with the addition of compatibilizer in untreated BP based composites in comparison to composites without compatibilizer with respective composition. When a sudden impact load is applied at the notch area more fiber pullout has been observed in case of untreated PP/BP composites, due to weak interfacial adhesion between the fiber and the matrix.

There is decrease in the izod impact in case of PP/BP (treated) composites also observed. The interfacial adhesion is improved in case of PP/BP (treated) composites as fiber pullout requires higher energy than fiber fracture. This interfacial adhesion is improved by the surface modification of the fibers which is also supported by the SEM micrographs [42].

3.2.3 Heat Deflection Temperature (HDT)

Table 4 shows that the HDT of the composites increase with increasing reinforcement contents in them. The composites processed with treated reinforcements is further higher than those of the composites processed with untreated reinforcements except PP40BE-4 and PP40BE-6. The increase in HDT confirms increase in the thermal stability of the composites and therefore can be used for higher temperature applications. The loading of fiber in matrix resulted in increase in HDT which could be due to increase in restricted mobility of polymer chains. Increase in HDT has been reported by Bispo et al with the incorporation of *Curaua fiber* in polypropylene composites [43]. With the addition of compatibilizer, a further increase in HDT was observed which could be due to good interfacial strength [44]. The chemical treatments of fibers have led to increase in the HDT of PP/BP composites in comparison to enzyme treatment

Accepted Manuscript

of fibers. With increase in enzyme treatment time HDT of the composites decreases in comparison to untreated compatibilized PP/BP composite.

Table 3. Mechanical Properties of PP/BP composites with and without compatibilizer

Sample Designation	Tensile Strength at yield (MPa)	Tensile Modulus (MPa)	% Elongation at yield	Flexural strength (MPa)	Flexural Modulus (MPa)	Impact Strength (J/m)
PP	33.0 ± 0.2	815 ± 1.3	9.9 ± 0.03	37.9 ± 0.7	1199 ± 5.5	21 ± 0.8
PP10B	27.8 ± 1.8	830 ± 105	5.8 ± 0.04	36.0± 0.4	1125 ± 35.3	28 ± 2.3
PP20B	26.3 ± 1.8	854 ± 136	5.4 ± 0.04	38.8 ± 2.1	1332 ± 31.2	29 ± 1.1
PP30B	24.0 ± 1.5	1128 ± 50	4.2± 0.2	38.2 ± 0.6	1391 ± 32.7	30 ± 1.5
PP40B	22.2± 1.5	1378 ± 24.5	3.6± 0.17	39.0 ± 0.6	1622 ± 64.4	32 ± 1.3
PP20B1C	26.5 ± 0.1	899 ± 50	7.0 ± 0.07	38.0 ± 1.1	1432 ± 45.7	21 ± 1.4
PP20B2C	26.8 ± 0.2	896 ± 21.4	7.1 ± 0.1	36.5 ± 0.5	1294 ± 49.3	22 ± 1.5
PP20B4C	27.0 ± 0.2	899 ± 11.0	7.0 ± 0.19	38.0 ± 0.2	1274 ± 8.3	22 ± 1.0
PP30B1C	31.0 ± 0.3	1524 ± 117	4.2 ± 0.15	37.5± 0.7	1346 ± 13.1	24.3 ± 0.9
PP40B1C	32.8± 0.9	1585± 83	3.3 ± 0.28	39.3± 0.8	1722± 46.2	24± 0.01
PP30BA	33.4 ± 2.9	1600 ± 162	3.6 ± 0.2	44.8 ± 1.4	1744 ± 49.3	19.4 ± 3.3
PP30BK	31.4± 1.0	1587± 88	3.5 ± 0.5	43.9 ± 1.8	2055 ± 121	23 ± 1.3
PP30BS	33.5± 0.5	1817± 75	3.8± 0.2	47.4 ± 0.6	1688± 39	18 ± 1.5
PP30BE-2	30.9 ± 0.5	1999 ± 80	1.9 ± 0.05	45.4 ± 1.1	1892 ± 75.6	18 ± 0.5
PP30BE-4	30.2 ± 0.4	1645 ± 21.4	3.4 ± 0.11	46.7 ± 0.7	1829 ± 72	19 ± 0.95
PP30BE-6	32.3 ± 1.3	1688 ± 11.0	3.6 ± 0.14	45.9 ± 0.9	1794 ± 33.2	20 ± 1.3
PP40BA	33.0 ± 0.3	2025 ± 42	3.6 ± 0.2	44.6 ± 2.5	2025 ± 81.3	19.2± 2.0
PP40BK	30.3 ± 1.0	1956 ± 56	3.4 ± 0.4	45± 1.7	1753 ± 54.0	22.5 ± 1.3
PP40BS	35± 0.3	2106 ± 95	3.3 ± 0.1	51± 0.5	2043 ± 194.8	20 ± 0.82
PP40BE-2	31 ± 0.5	2011 ± 50	2.6 ± 0.08	45.3 ± 1.7	2256 ± 75.7	19.2 ± 0.5
PP40BE-4	30.1 ± 0.4	1590 ± 21.4	3.2 ± 0.3	44.6 ± 1.4	2083 ± 78.9	20.5 ± 1.4
PP40BE-6	28.2 ± 1.0	1412 ± 27.8	3.4 ± 0.05	40 ± 2.2	1771.5 ± 70.3	20 ± 3.0

Accepted Manuscript

Table 4. HDT of PP/BP composites

Sample	HDT (ºC)	Sample	HDT (ºC)
PP	100	PP30BK	126
PP10B	105	PP30BS	127
PP20B	108	PP30BE-2	118
PP30B	115	PP30BE-4	113
PP40B	122	PP30BE-6	94
PP20B1C	113	PP40BA	132
PP20B2C	116	PP40BK	137
PP20B4C	115	PP40BS	136
PP30B1C	120	PP40BE-2	125
PP40B1C	128	PP40BE-4	119
PP30BA	125	PP40BE-6	96

3.2.4 Dynamic mechanical analysis of PP/BP composites

Storage modulus of PP/BP composites

The DMA graph and calculated values for storage modulus are given in Figures 4 & 5 and Table 5 respectively. The increase in storage modulus on adding untreated BP fiber with maximum increase in case of PP40B in comparison to PP has been achieved. This could be due to the reinforcement imparted by the BP fiber which allows stress transfer from the matrix to the fiber [45]. For example, at 25°C, the G′ value for neat PP is 1771 MPa and for PP40B is 2385 MPa, which indicates ~ 35% increase in stiffness. A sudden decrease in the storage modulus is observed in the temperature region -15°C to +15°C (Figure 4) and the rate of decrease is lessened when the polymer reaches the rubbery zone. In the glassy zone, contribution of the fiber stiffness to the material modulus is minimal [46]. In case of PP neat, a sudden fall in the storage modulus is observed when the transition takes place from glassy to rubbery zone in comparison to PP/BP composites due to increased mobility of polymer chains above Tg. Further on addition of compatibilizer, storage modulus increases as can be seen with PP/BP composite PP20B1C in comparison to PP20B composite. This confirms an enhancement in the stiffness of the matrix with better adhesion or interaction. The storage modulus enhancement was comparatively more in case of PP20B1C than PP20B2C and PP20B4C as observed in Figure 4b.

Accepted Manuscript

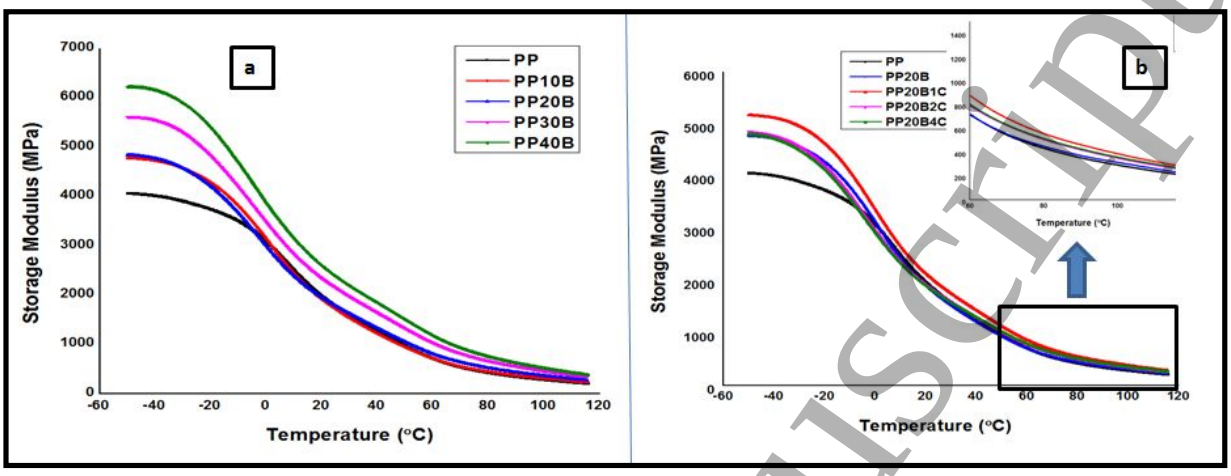


Figure 4. Storage Modulus curve for (a) untreated and (b) untreated compatibilized PP/BP composites

Table 5. Storage Modulus data at different temperature for PP/BP composites

Sample	E' (-40ºC)	E' (25ºC)	E' (115ºC)
PP	4032	1771	222
PP10B	4739	1712	240
PP20B	4787	1782	279
PP30B	5566	2150	351
PP40B	6178	2385	393
PP20B1C	5133	1944	294
PP20B2C	4792	1754	270
PP20B4C	4742	1741	276

The chemical treatment of the BP fiber also aided in the increase in the viscoelastic properties of the composite. This is in accordance with other similar studies [45,47]. Figure 5 shows the effect of chemical treatments of fibers on the storage modulus of the composites at 30% and 40% fiber loading at a frequency of 1 Hz. The PP/BP (treated) composites have a higher modulus than the PP/BP (untreated) composite which confirms better interfacial adhesion between fiber and matrix after the treatment leading to less molecular mobility in the interfacial adhesion [45] and increasing the interfacial adhesion and compatibility of BP fiber and the PP

Accepted Manuscript

matrix [48]. From the storage modulus curves it is observed that alkali treatment of fiber has led to the highest increase in the reinforcement effect on the PP/BP composites

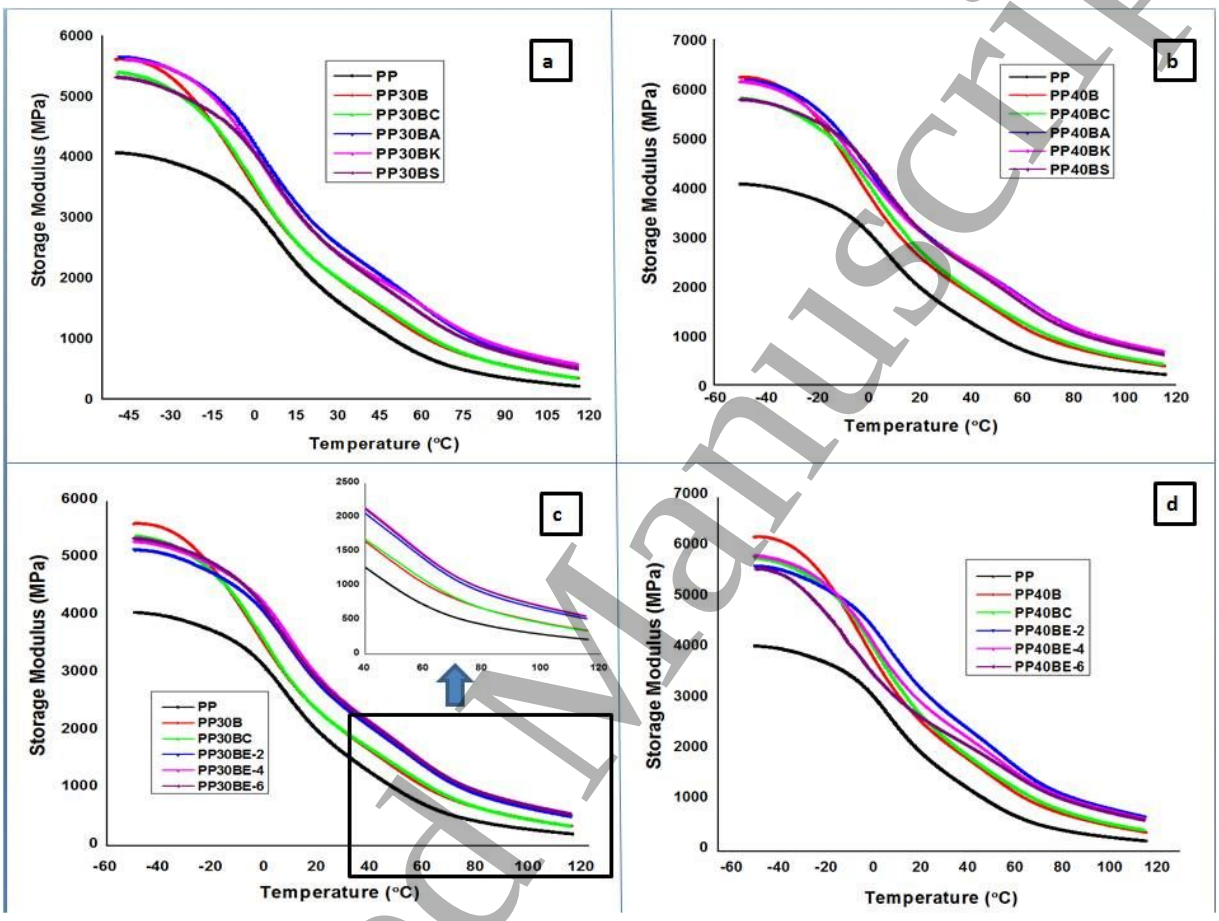


Figure 5. Storage Modulus (a) 30wt% PP/BP chemical treated composites (c) 30wt% PP/BP enzyme treated composites and 30wt% PP/BP untreated composites (b) 40wt% PP/BP chemical treated composites and (d) 40wt% PP/BP enzyme treated composites and 40wt% PP/BP untreated composites

The percentage increase of storage modulus of the treated PP/BP composites at a 25ºC temperature in comparison to PP matrix at same temperature has been tabulated in Table 6.

Accepted Manuscript

Table 6. Percent increase in stiffness of PP/BP composites w.r.t to PP

Sample	% Increase in stiffness at 25º C wrt PP	Sample	% Increase in stiffness at 25º C wrt PP
PP30B	21%	PP40B	35%
PP30BC	23%	PP40BC	41%
PP30BA	53%	PP40BA	66%
PP30BK	46%	PP40BK	67%
PP30BS	45%	PP40BS	64%
PP30BE-2	46%	PP40BE-2	70%
PP30BE-4	52%	PP40BE-4	56%
PP30BE-6	49%	PP40BE-6	42%

3.2.5 Loss Modulus of PP/BP composites

The Figure S9 and S10 shows that the fiber loading in matrix lowers the Tg. This means that transition from glassy to rubbery region can occur at lower temperature which can be concluded that the BP are increasing the mobility of PP chains [49]. With the increase in temperature a shoulder is observed which gets more noticeable on addition of BP fiber (Figure S9). The use of treated fibers led to the shift in the glass transitions temperature of the PP/BP composites. A sharp peak at ~6.5°C in PP indicated the Tg relaxations corresponding to amorphous phase. A shoulder is observed in the temperature range of 45-85°C, on the addition of the BP fiber to PP matrix. A shift of the temperature towards the higher side is observed with broader transition zone in treated PP/BP composites (Figure S10). The shift in the Tg to higher temperature could be due to decreased mobility of the polymer chains due to the hindrance provided by the reinforcement effect of natural fiber in the matrix [50]. The loss modulus of untreated and treated PP/BP composites are given in Table S4.

3.2.6 Tan Delta of PP/BP composites

Tan delta measures the dissipation of a material. Systems with higher tan δ peaks have higher ratios of energy absorbing viscous motions and are therefore generally tougher systems than those with low tan δ amplitudes [50]. As the fiber concentration in the matrix is enhanced,

Accepted Manuscript

the tan δ peak becomes broad and the area under the peak reduces (Figure S11) as the polymer fraction in the composite is lessened and only the amorphous phase of the partially crystalline PP is involved in glass transition.

The effect of treatment of fiber on the damping effect of the PP/BP composites was also studied. Figure S12 depicts the two relaxation peaks located between 4-12°C (Tα) and 75-100°C ($T\alpha^*$) respectively and the values have been reported as Table S5. The Tα relaxations represents the glass transition and Tα* is attributed to the lamellar slip and rotation in the crystalline phase. It has been observed that the damping value decreases for the treated fibers composites than the untreated fiber composites and in some cases the value is less than PP. Higher the tan δ peak more is the mobility of the polymeric chains. A decrease in damping is also supported by the literature [51] which indicates good interface bonding.

3.2.7 Coefficient of Reinforcement (C) of PP/BP composites

The effectiveness of the filler is inversely proportional to the value of C i.e. the greater the value of the constant C, the lower the effectiveness of the filler. The measured E' values at -40°C (E'_G) and 90°C (E'_R) for PP and the composites were employed. The value of 'C' decreases with the treatment of fiber (Table S6). This indicates that maximum stress transfer between the fiber and the matrix takes place at this fiber loading which is also supported by the similar study [52].

3.2.8 Adhesion study of PP/BP composites

Adhesion factor is used for the verification of the effect of treatments on the interfacial adhesion between BP fibers and PP matrix [53]. Figure S13 depicts low adhesion factor curves of treated fiber compared to untreated PP/BP composites which confirms the enhancement in fiber matrix adhesion due to fiber treatments.

3.3 Thermogravimetric analysis of PP/BP composites

TG/DTG traces for PP and PP/BP composites in absence and presence of compatibilizer [PPB and PPBC] are shown in Figures 6 and 7. From the scan of TGA and DTG curves it is clear that the decomposition pattern for PP and PP/BP composites is different as PP degradation is single step whereas PP/BP composites is a three-step degradation at around ~278º C, 450ºC

Accepted Manuscript

and 745ºC. As these fibers have strong affinity towards water, the first weight loss step in the TGA scan corresponds to the weight loss of water. The degradation or weight loss in the temperature range of 250-320ºC and 350-650ºC conform to loss of cellulose - hemicelluloses and lignin respectively. In case of PP, the thermal degradation started at around 429 ºC and the T_{max} was found to be 450ºC. Two step degradation is observed in composites with peak degradation temperatures. With the incorporation of BP fibers in PP matrix the onset degradation of temperature of composites increased by ~6ºC (PP40B).

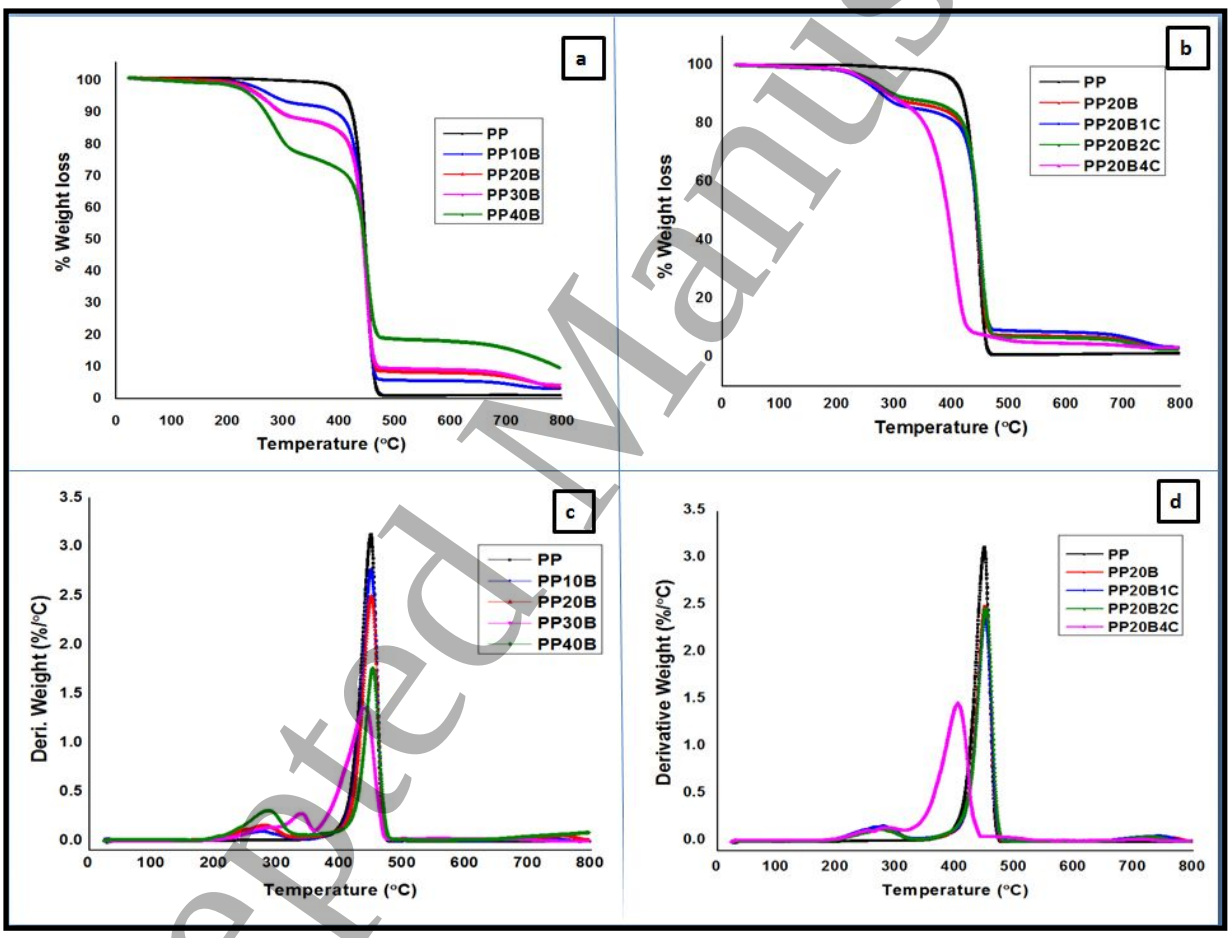


Figure 6. (a) % Weight loss and b) Deri. Weight for untreated PP/BP composites (c) % Weight loss and (d) Deri. Weight for compatibilized untreated PP/BP composites

A slight decrease in the onset degradation temperature is observed with the incorporation of compatibilizer. After treatment of fibers it has been observed that the thermal stability of the PP/BP composites have been increased 5ºC in comparison to BP fiber. Peak degradation

Accepted Manuscript

temperature in case of alkali treated fiber composites is more which depicts the increase in thermal stability of the composite in comparison to untreated fiber composite of same fiber loading. No peak degradation is observed at around 750 ºC which is in treated composites which further confirms that lignin has been removed after treatment of fiber. A char yield of 3-11% is observed in treated PP/BP composites as depicted in Table 7.

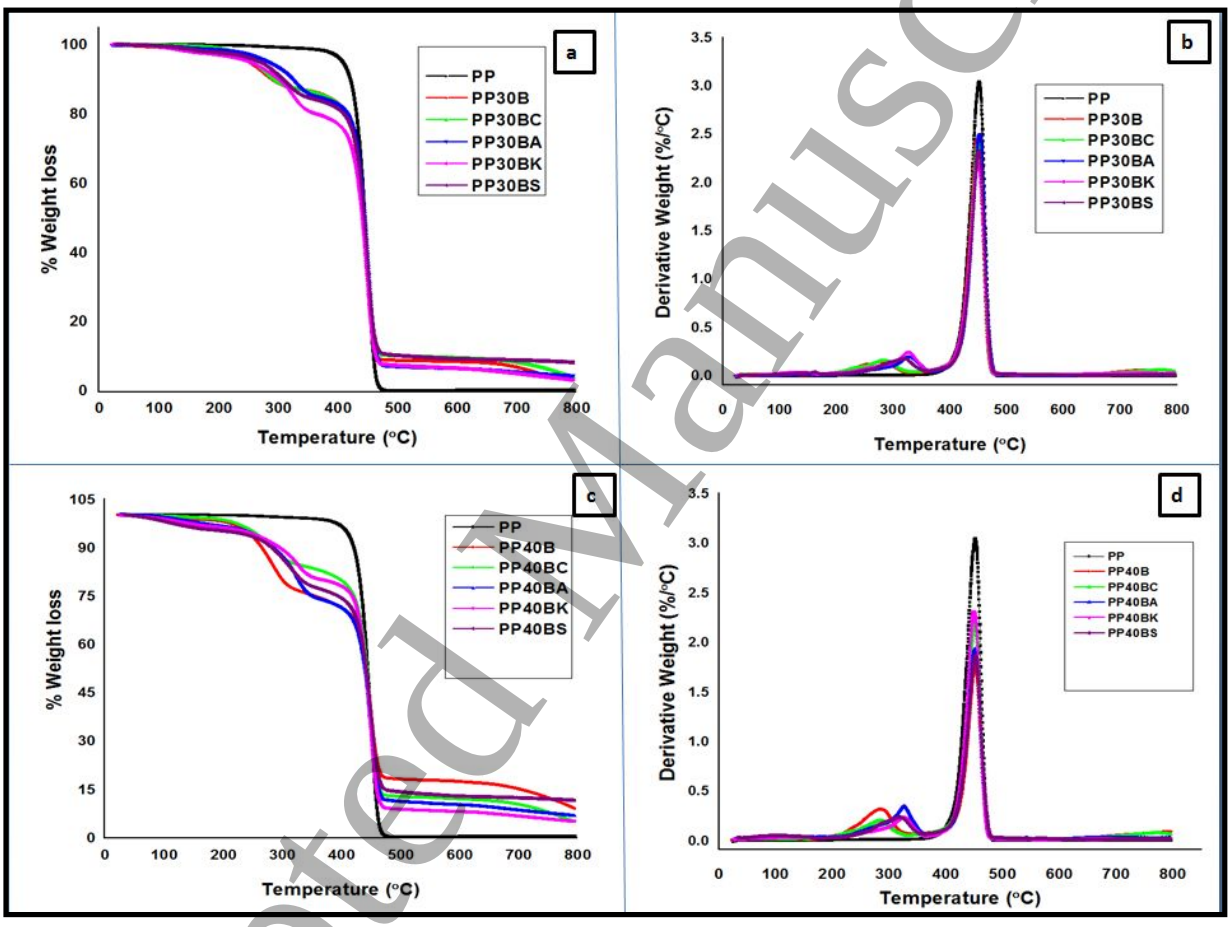


Figure 7. (a) % Weight loss and (b) Deri. Weight for 30wt % chemical treated PP/BP composites and 30wt% PP/BP untreated composites (c) % Weight loss and (d) Deri. Weight for 40wt % chemical treated PP/BP composites and 40wt% PP/BP untreated composites

3.4 X-Ray Diffraction (XRD) for PP/BP composites

Figures S14 and S15 shows the wide-angle X-ray diffraction patterns of untreated and treated PP/BP composites. Five peaks in the 2θ range of 10–30˚ at (110), (040), (130), (111) and

Accepted Manuscript

(041) lattice plane in XRD scan for PP were observed which confirms the monoclinic α form without any peculiar peak corresponding to beta (β) and gamma (γ) form of PP.

Table 7. The onset, peak degradation and % char residue of PP/BP composites

Sample	Onset Degradation temperature (ºC)	Peak Degradation temperature (ºC)	% Char or Residue left
PP	430	450	0.31
PP-g-MA	380	429	0.54
PP10B	236, 429	277, 449	1.35
PP20B	229, 440	285, 406	2.29
PP30B	240, 432	285, 451	5.2
PP40B	238, 435	287, 453	8.27
PP20B1C	225, 431	278, 450	3.35
PP20B2C	226, 431	275, 451	3.46
PP20B4C	230, 399	279, 430	3.73
PP30B1C	231, 430	280, 450	3.3
PP40B1C	253, 432	287.8, 450	4.7
PP30BA	285, 433	382, 452	5.2
PP30BK	284, 431	327, 430	3.0
PP30BS	269, 432	320, 451	8.2
PP30BE-2	278, 407	325, 442	4.3
PP30BE-4	281, 402	326, 440	4.2
PP30BE-6	280, 433	328, 452	6.3
PP40BA	285, 432	326, 451	4.6
PP40BK	281, 432	325, 450	8.9
PP40BS	271, 433	319, 452	11.2
PP40BE-2	280, 435	325, 454	8.0
PP40BE-4	279, 435	327, 454	7.9
PP40BE-6	280, 391	326, 419	5.5

Accepted Manuscript

A peak that conforms to beta crystalline phase at 2 θ =16.30 (degrees) represents the (300) diffraction plane in PP/ BP composites. On incorporation of PP-g-MA, the two forms of crystals still existed in PP/BP composites.

With the addition of compatibilizer the K value is observed to decrease (Table 8). The lower K-values implies that the addition of PP-g-MA significantly decreases the β-form crystal content in the composite. These results are consistent with the previous study [54]. In chemical treated fibers, a drop in K value is observed whereas in enzyme treatment the K value increases thereby confirming an increasing in beta phase of PP.

Table 8. k values for PP and PP/BP composites

Samples	K value	Samples	K value
PP	Nil	PP20B1C	0.14
PP10B	0.19	PP20B2C	0.10
PP20B	0.15	PP20B4C	0.07
PP30B	0.10	PP30BC	0.15
PP40B	0.09	PP40BC	0.16
PP30BA	0.25	PP40BA	0.20
PP30BK	0.19	PP40BK	0.10
PP30BS	0.27	PP40BS	0.14
PP30BE-2	0.26	PP40BE-2	0.22
PP30BE-4	0.15	PP40BE-4	0.18
PP30BE-6	0.12	PP40BE-6	0.17

3.5 Crystallization behavior study of PP/BP composites

DSC data of PP, PP-g-MA, and PP/BP composites without /with compatibilizer have been summarized in Table S7. The melting temperature (T_m) of samples were taken from the second heating scan. There was no change in the T_m but a slight drop in Tc for sample PP40B where Tc and Tm remained almost same on incorporation of compatibilizer. This indicates that addition of compatibilizer does not seem to affect the close packing of PP/BP composites. From the area under the endothermic transition, the heat of fusion was determined which was used to calculate percent crystallinity.

Accepted Manuscript

An exothermic transition due to crystallization was observed in the cooling scans of all the samples. Percent crystallinity showed an increase for samples PP10B, PP20B and PP30B. On the other hand, there is no change in the crystallinity of compatibilized composites. In case of some of the treated PP/BP composites (Table S7) an increase in Tc and % crystallinity is observed in alkali, silane and enzyme treated composites. 13% increase in crystallinity is observed in case of PP30BA. This indicates that treated BP fiber is acting as a nucleating agent which enhances the crystallization rate and thus show the nucleation effect [55]. The treatments of BP fiber are expected to remove the amorphous parts in the natural fiber and therefore the crystalline parts increase after treatments [56].

3.6 Water Absorption study of PP/BP composites

Water Absorption of PP is almost nil as compared to PP/BP composites as observed from the Figures S16 and S17. On addition of the natural filler from 10-40wt% (BP fiber) to the PP the water absorption of the composite (PP10B to PP40B) increased in comparison to PP. This phenomenon has also been cited in the work [57] which could be due to hydrophilic nature of BP fiber which lead to water diffusion in composites due to the presence of micro voids and less interfacial adhesion between fiber and the matrix. But with the addition of the compatibilizer (PP-g-MA) to the composites (PP20B-1C to 4C), it decreases the water absorption capacity in comparison to PP20B (Figure S16) which therefore confirms the better interfacial adhesion between matrix and fiber and the reduction of micro voids in the composite as the polar hydroxyl group of fiber is no longer available due to formation of covalent bonds of anhydride group in PP-g-MA and hydroxyl groups of cellulose. For the composites prepared from the treated BP fibers (alkali, potassium permanganate, silane and enzyme treated) the water absorption rate decreases (Figure S17). The decrease is observed more in the silane treated composite The same phenomena have been observed in the study [58] as the treatment of fibers leads to the modifications which thereby reduce the hydrophilic nature of the BP fiber so enhancing the interfacial adhesion between the fiber and matrix.

4. Conclusion

The effect of banana peel fibrous powder (BP) content, treated and untreated both, on the morphological, mechanical, visco-elastic behavior and thermal properties of PP has been

Accepted Manuscript

achieved. This established potential of agricultural waste to be used as reinforcing agents. The pre–treatments of fibers aided in good dispersion of fibers in case of composites prepared from treated BP fiber. The enhancement in the mechanical properties and storage modulus on adding BP fibers confirmed reinforcing ability of BP fibers on PP matrix. Addition of compatibilizer also aided to the increase in the properties. The PP/BP composite with treated BP showed better thermal stability. The HDT results also revealed an increase in thermal stability on loading of fiber to the matrix. The increase in HDT observed for treated PP/BP composites than the untreated PP/BP composites. The XRD data revealed an increase in crystallinity on treatment of fibers in comparison to untreated BP fiber. With the incorporation of fibers, a beta phase in PP was observed which decreases on addition of compatibilizer and in case of treated PP/BP composites when compared with untreated PP/BP composites. The addition of BP fiber in PP matrix led to increase in water absorption of PP/BP composites which declined on addition of compatibilizer. A further decline in water absorption was observed for treated PP/BP composites. This confirms that these can be used for outdoor applications.

Acknowledgements

The authors would like to acknowledge Analytical Department of Indian Oil Corporation Ltd (IOCL) Research and Development Centre, Faridabad, India for their support and service provided. The authors are grateful to Delhi Technological University for providing necessary facilities to carry out this work.

Supplementary data related to essential information is submitted along with this article.

References

[1] Koronis G, Silva A and Fontul M 2012 Green composites: A review of adequate materials for automotive applications *Composites: Part B* **44** 120-127.

[2] Zykova A K, Pantyukhov P V, Monakhova T V and Popov A A 2017 Degradation kinetics of ethylene-octene copolymer/wood flour biocomposites in dependence to filler content *IOP Conf Ser: Mater Sci Eng* **213** 012012.

Accepted Manuscript

[3] Al-Maharma A Y and Sendur P 2019 Review of the main factors controlling the fracture toughness and impact strength properties of natural composites *Mater Res Express* **6** 022001.

[4] Rahmaniar and Susanto T 2019 Impacts of rice husk ash filler loading on curing, morphological characteristics and tensile properties of natural rubber/ethylene propylene rubber blends *IOP Conf Ser: Mater Sci Eng* **509** 012116.

[5] Singh T, Gangil B, Patnaik A, Kumar S, Rishiraj A and Fekete G 2019 Physico-mechanical, thermal and dynamic mechanical behaviour of natural-synthetic fiber reinforced vinyl ester based homogenous and functionally graded composites *Mater Res Express* **6** 025704.

[6] Mittal M and Chaudhary R 2019 Biodegradability and mechanical properties of pineapple leaf/coir Fiber reinforced hybrid epoxy composites *Mater Res Express* **6** 045301.

[7] Luthra P, Singh R and Kapur G S 2019 Preparation and studies of Pigeon Pea Stalk/ Polypropylene Composites with and without Compatibilizer *Polym Polym Compos* in press.

[8] Kulkarni M B, Radhakrishnan S, Samarth N and Mahanwar P A 2019 Structure, mechanical and thermal properties of polypropylene based hybrid composites with banana fiber and fly ash *Mater Res Express* **6** 075318.

[9] AlMaadeed M A, Kahraman R, Noorunnisa K P and Al-Maadeed S 2013 Characterization of untreated and treated male and female date palm leaves *Mater Des* **43** 526–531.

[10] Luthra P, Singh R and Kapur G S 2019 Effect of chemical treatments on Pigeon Pea stalk fibers *Eur J Sci* in press.

[11] Vijay R, Vinod A, Kathiravan R, Siengchin S and Singaravel D L 2018 Evaluation of *Azadirachta indica* seed/spent *Camellia sinensis* bio-filler based jute fabrics–epoxy composites: Experimental and numerical studies *J Indust Textil* 1–26.

[12] Ganeshan P, NagarajaGanesh B, Ramshankar P and Raja K 2018 Calotropis gigantea fibers: A potential reinforcement for polymer matrices *Int J Polym Anal Character* **23** 271-277.

Accepted Manuscript

[13] Majid R A, Mohamad Z, Rusman R, Zulkornain A A, Halim N A, Abdullah M, Low J H, 2018 Development of tea waste/kapok fiber composite paper *Chem Eng Trans* **63** 457-462.

[14] Gayatri U, Malkapuram R, Vasu AT and Chavali M 2016 Fabrication and testing of composite powder material from prawns shells *RRJOMS* **4** 37-46.

[15] Nijandhan K, Muralikannan R and Venkatachalam S 2018 *Ricinus communis* fiber as potential reinforcement for lightweight polymer composites *Mater Res Express* **5** 095307.

[16] Dungani R, Karina M, Sulaeman S A, Hermawan D and Hadiyane A 2016 Agricultural waste fibers towards sustainability and advanced utilization: A review *Asian J Plant Sci* **15** 42-55

[17] Ramli N, Mazlan N, Ando Y, Leman Z, Abdan K, Aziz A A and Sairy N A 2018 Natural fiber for green technology in automotive industry Asian *J Plant Sci IOP Conf Series: Mater Sci Eng* **368**

[18] Khan H 2017 Top 10 largest banana producing state in India. *https://www.worldblaze.in/largest-banana-producing-state-in-india/*. Accessed 1 December 2018.

[19] Orozco R S, Hernóndez P B, Morales G R, Núñez F U, Villafuerte J O, Lugo V L, Ramíez N F, Díaz C E B and Vózquez P C 2014 Characterization of lignocellulosic fruit waste as an alternative feedstock for bioethanol production *BioRes* **9** 1873-1885

[20] Pereira P H F, Benini K C C C, Watashi C Y, Voorwald H J C and Cioffi M O H 2013 Characterization of high-density polyethylene reinforced with banana peel fibers *BioRes* **8** 2351-2365

[21] Naidu A L, Raghuveer D and Suman P 2013 Studies on characterization and mechanical behavior of banana peel reinforced epoxy composites *Int. J. Sci. Eng. Res.* **4** 844-851

[22] Akbay I K and Gungor A 2018 Investigation of using Banana peels in EPDM as bio-based filler *Int. Adv. Res Eng. J.* **2** 132-136.

[23] Zhang X, Wang F and Keer L M 2015 Influence of surface modification on the microstructure and thermo-mechanical properties of bamboo fibers *Materials* **8** 6597–6608.

[24] Cao Y, Chan F, Chui Y H and Xiao H 2012 Characterization of flax fibers modified by alkaline, enzyme, and steam heat treatments *BioRes.* **7** 4109-4121.

Accepted Manuscript

[25] Memon J R, Memon S Q, Bhanger M I, Memon G Z, Turki A E and Allen G C 2008 Characterization of banana peel by scanning electron microscopy and FT-IR spectroscopy and its use for cadmium removal *Colloids Surf. B* **66** 260–265

[26] Schwarzová I 2016 Investigation of observed changes in treated hemp hurds *GeoSci. Eng.* **62** 22-26.

[27] Li Y Y, Du G M, Feng X J, Mu Y W and Li D Q 2016 Extraction and hydrophobic modification of cotton stalk bark fiber *Int. J. Polym. Sci.* Article ID 8769360 6 pages.

[28] Zhang X, Wang F and Keer L M 2015 Influence of surface modification on the microstructure and thermo-mechanical properties of bamboo fibers *Materials (Basel)* **8** 6597-6608.

[29] Zhu J, Brington J, Zhu H and Abhyankar H 2015 Effect of alkali, esterification and silane surface treatments on properties of flax fibres *J. Sci. Res. Rep.* **4** 1-11

[30] Carvalho K C C, Mulinari D R, Voorwald H J C and Cioffi M O H 2010 Chemical modification effect on the mechanical properties of HIPS/Coconut fiber composites *BioRes*. **5** 1143-1145.

[31] Terinte N, Ibbett R and Schuster K C 2011 Overview on native cellulose and microcrystalline cellulose I structure studied by X ray diffraction (WAXD): Comparison between measurement techniques *Len. Ber.* **89** 118-131

[32] Anuar H, Hassan N A and Fauzey F M 2011 Compatibilized PP/EPDM-kenaf fibre composite using melt blending method *Adv. Mat. Res.* **264-265** 743-747

[33] Fatimah A R H, Supri A G and Firuz Z 2015 The effect of polyethylene-grafted-maleic anhydride as compatibilizer on properties of recycled high-density polyethylene/ethylene vinyl acetate/taro powder (colocasia esculenta) composites *J. Adv. Res. Mat. Sci.* **10** 26-39.

[34] Rahman Md R, Hamdan S, Hashim D M A, Islam Md S and Takagi H 2015 Bamboo fiber polypropylene composites: Effect of fiber treatment and nano clay on mechanical and thermal properties *J. Vinyl Addit. Technol.* **21** 253–258.

Accepted Manuscript

[35] Mohammed L, Ansari M N M, Pua G, Jawaid M and Islam MS 2015 A Review on natural fiber reinforced polymer composite and its applications *Int. J. Polym. Sci.* Article ID 243947, 15 pages.

[36] Sujaritjun W, Uawongsuwan P, Art W P and Hamada H 201) Mechanical property of surface modified natural fiber reinforced PLA biocomposites *Energy Proc*. **34** 664-672

[37] Chen X, Zhou L, Pan X, Hu J, Hu Y and Wei1 S 2016 Effect of different compatibilizers on the mechanical and thermal properties of starch/polypropylene blends *J. Appl. Polym. Sci.* **133** 43332

[38] Lee H S and Cho D 2008 Effect of natural fiber surface treatments on the interfacial and mechanical properties of henequen/polypropylene biocomposites *Macromol. Res.* **16** 411-417

[39] Islam M R, Beg M D H and Gupta A 2013 Characterization of laccase-treated kenaf fiber reinforced recycled polypropylene composites *BioRes.* **8** 3753-3770

[40] Salisu A A, Yakasai M Y and Aujara K M 2016 Physico-mechanical properties of chemically modified sisal fibre reinforced unsaturated polyester composites *Int. J. Chem. Mol. Eng.* **10** 81-88

[41] Bachtiar D, Sapuan SM, Hamdan MM 2010 Flexural properties of alkaline treated sugar palm fibre reinforced epoxy composites *Int. J. Auto Mech. Eng.* **1** 79-90

[42] Kalia S 2018 Lignocellulosic Composite Materials *Spriger series* **153**

[43] Bispo S J L, Júnior R C S F and de Aquino E M F 2015 Mechanical properties analysis of polypropylene biocomposites reinforced with curaua fiber *Mater. Res.* **18** 833-837

[44] Chattopadhyay S K, Khandal R K, Uppaluri R and Ghoshal A K 2011 Bamboo fiber reinforced polypropylene composites and their mechanical, thermal, and morphological properties *J. Appl. Polym. Sci.* **119** 1619–1626

[45] Zhai Z, Liu Z, Feng L and Liu S 2014 Interfacial adhesion of glass fibre reinforced polypropylene-maleic anhydride modified polypropylene copolymer composites J. *Reinforced Plastics Compos.* **33** 785-793

[46] Huda M S, Drzal L T, Misra M, Mohanty A K, Williams K and Mielewski F D F 2005 A Study on biocomposites from recycled newspaper fiber and poly(lactic acid) *Ind. Eng. Chem. Res.* **44** 5593-5601

Accepted Manuscript

[47] Saba N, Jawaid M, Alothman O Y and Paridah M T 2016 A review on dynamic mechanical properties of natural fibre reinforced polymer composites *Constr. Build. Mater.* **106** 149-159

[48] Shah H, Srinivasulu B and Shit S C 2013 Influence of banana fibre chemical modification on the mechanical and morphological properties of woven banana fabric/unsaturated polyester resin composites *Polym. Renew. Res.* **4** 61-84

[49] Mofokeng J P, Luyt A S, Tabi T, Kovacs J 2011 Comparison of injection moulded, natural fibre-reinforced composites with PP and PLA as matrices *J. Thermoplast Comp. Mater.* **25** 927–948

[50] Karaduman Y 2018 Experimental investigation on the flexural and dynamic mechanical properties of jute fiber/cork-reinforced polyester sandwich composites *BioRes.* **13** 3958-3970

[51] Cinquin J, Chabert B, Chauchard J, Morel E and Trotignon J D 1990 Characterization of a thermoplastic (polyamide 66) reinforced with unidirectional glass fibers. Matrix additives and fibers surface treatment influence on the mechanical and viscoelastic properties *Composites* **21** 141-147

[52] Idicula M Malhotra S K Joseph K and Thomas S 2005 Dynamic mechanical analysis of randomly oriented intimately mixed short banana/sisal hybrid fibre reinforced polyester composites *Compos. Sci. Technol.* **65** 1077–1087

[53] Pakravan H, Jamshidi M and Masoud L 2012 Adhesion of polypropylene fiber to cement matrix *J. Adhesion Sci. Tech.* **26** 1383–1393

[54] Niu P, Liu B, Wei X, Wang X and Yang J 2010 Study on mechanical properties and thermal stability of polypropylene/hemp fiber composites *J. Reinfor. Plastics Comp.* **30** 36-44

[55] Lopez J P, Gironés J, Méndez J A, Mansouri N-E El, Llop M, Mutjé P and Vilaseca F 2012 Stone- ground wood pulp-reinforced polypropylene composites: water uptake and thermal properties *BioRes.* **7** 5478-5487

[56] Jose T E, Joseph A, Skrifvars M, Thomas S and Joseph K 2010 Thermal and crystallization behavior of cotton—polypropylene commingled composite systems *Polym. Compos.* **31** 1487-1494

Accepted Manuscript

[57] Kada D, Migneault S, Tabak G and Koubaa A 2016 Physical and mechanical properties of polypropylene- wood – carbon fiber hybrid composites *BioRes.* **11** 1393-1406

[58] Meenalochani K S and Reddy B G V 2017 A Review on water absorption behavior and its effect on mechanical properties of natural fibre reinforced composites *Int. J. Innovat. Res. Adv. Eng* **4** 143-147

*****

Accepted Manuscript

AIP Conference Proceedings

Ecosystem services by wetlands for polluted water treatment

Cite as: AIP Conference Proceedings **2121**, 160002 (2019); https://doi.org/10.1063/1.5115973
Published Online: 18 July 2019

Padma Vasudevan, P. K. Sen, R. K. Srivastava, Mamata Tandon, Shriram Hegde, T. Vijaya Kumar, and Philip Davies

View Online
Export Citation

AIP Conference Proceedings
Get 30% off all print proceedings!
Enter Promotion Code PDF30 at checkout

AIP Publishing

© 2019 Author(s).

Ecosystem Services by Wetlands for Polluted Water Treatment

Padma Vasudevan[1.a)], P.K. Sen[1,b)] R.K. Srivastava[2,c)] , Mamata Tandon[1,d)], Shriram Hegde[1,e)], T.Vijaya Kumar[3,f)] ,and Philip Davies[4,g)]

[1]*Indian Institute of Technology Delhi, India,*
[2]*G. B. Pant University of Agriculture and Technology, Pantnagar, India,*
[3]*Delhi Technological University, Delhi, India* [4]*Birmingham University, U.K.*

[a)]Corresponding author: padmav10@gmail.com
[b)]pksen45@gmail.com
[c)]rajeevsrivastava08@gmail.com
[d)]mamta.tandon@gmail.com
e)hedge@am.iited.ac.in
[f)]tvijayakumar1970@gmail.com
[g)]P.A.Davies@bham.ac.uk

Abstract. Various designs are discussed for ecosystem services in engineered wetlands for contaminated water treatment. Techniques such as Rhizo filtration, Constructed wetlands and Floating rafts are highlighted based on our experiments. Specifically, enhanced bioenergy production by domestic wastewater treatment in rhizo filtration by fast growing trees is discussed. In another set of experiments, designs and factors which can enhance the rate of pollutant removal from wastewater by floating wetlands are seen. In a study extending over three years fast growing short rotation trees, eucalyptus (Eucalyptus hybrid), poplar (Populus deltoides), salix (Salix alba) and melia (Melia azedarach) were irrigated with domestic wastewater with ground water as control. It was seen that the calorific value of woody biomass grown using ground water was of the order 17.4-15.6MJ/kg and this was increased by 6.7 to 16% on wastewater irrigation. Thus, besides a significant increase in the yield biomass (2-4 times) there is a concurrent increase in the calorific value of wood, resulting in higher energy output by wastewater irrigation. TGA and FTIR of the samples did not show significant difference between waste water and ground water irrigated biomass. In the floating wetland study Canna indica was taken on floating rafts and the phytoremediation efficiency of the units was examined. Further the floating raft units were integrated with a constructed wetland to form a hybrid system to enhance the overall efficiency in water treatment. Changes in waste water characteristics with contact time were examined. Reduction in BOD and COD of waste water significant in 2-5 days of contact and this varies with the maturity and density of plants and roots. There is a significant reduction in bacterial load in terms of E. coli counts in parallel with BOD. Rate of reduction of contaminants is near first order. The dissolved oxygen (DO) levels increase with time.

INTRODUCTION

Globally, nine planetary boundaries are considered collectively as limits, crossing the thresholds of which can be catastrophic. These are Stratospheric ozone depletion, Loss of biosphere integrity (biodiversity loss and extinctions), Chemical pollution and the release of novel entities, Climate Change, Ocean acidification, Freshwater consumption and the global hydrological cycle, Land system change, and Nitrogen and phosphorus flows to the biosphere and oceans. We can develop within the limits by emulating nature which is constantly recycling matter and energy in natural cycles within the ecological systems. Some strategies are to create *closed water cycles*, *closed energy cycles* and adopt *circular economy* and design *decentralized treatment systems* at the right scales.

Phytoremediation: This may be defined as 'the engineered use of plants for environmental remediation of water and soil. Selected plant species have a high potential to remove pollutants. Further a variety of micro flora in the water and soil - participate in this complex system which is useful in remediating industrial and domestic and agricultural waste waters.

8th BSME International Conference on Thermal Engineering
AIP Conf. Proc. 2121, 160002-1–160002-11; https://doi.org/10.1063/1.5115973
Published by AIP Publishing. 978-0-7354-1861-5/$30.00

In Constructed Wetlands (CW) which act as bio fillers, the wastewater passes slowly through shallow areas of dense aquatic vegetation which includes selected plants on a substrate (gravel, soil etc.), which hold the plants. The rate of flow and residence time is controlled by design and treated water with reduced pollutant load is collected at the outlet. CWs are classified according to the water flow regime as: Horizontal flow, Vertical flow and free surface constructed wetlands. Combined flow regimes, in so called hybrid constructed wetlands, exploit the specific advantages of the different systems selected and integrated.

Vegetative Filters, or Rhizofilters are plant-based natural treatment systems, combining fertilizer inputs with irrigation (Fertigation). This principally involves phytoremediation strategies for treatment of wastewater by irrigating fast-growing woody trees, green belts including grasses, amenity trees, and/or herbaceous perennials. In particular, dendro remediation/vegetative filters which use tree-based pollutant sequestration systems provide enhanced ecosystem services as they offset CO_2 emissions, generate oxygen and renewable energy in the form of fuel-wood, and perpetuate the nutrient cycles – thereby bringing economic returns and other benefits to society as a whole.

Floating wetlands or floating raft treatment: In this treatment system the waste water flows under floating rafts on which the plants grow. This is an eco-technological model involving aquatic macrophytes and various microorganisms colonized on their root hair as bio-films, which together remediate wastewater. Aquatic macrophytes are key components and play a crucial role in pollutant removal through rhizo filtration, rhizo degradation, root induced chemical and bacterial mechanisms, phyto accumulation etc, thus promoting efficient functioning. The rate of removal of pollutants follows near first order kinetics and depends on plant species especially the root system, the pollution load, the rate of flow of water.

In this paper, salient results of our study on two of the above systems designed and tested by us, are presented. These are 1. Rhizo Filtration and Fertigation experiments with selected fast growing tree species, and 2. Floating wetlands trials at pilot and field level as standalone and in integrated hybrid systems.

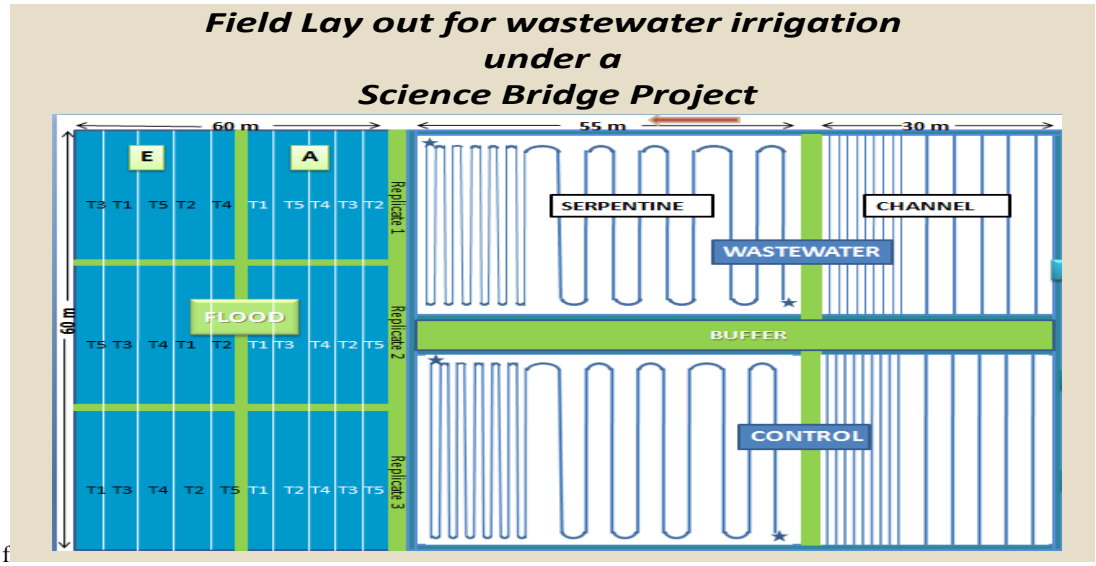
1. RHIZO FILTRATION: FERTIGATION FOR ENHANCED BIOENERGY PRODUCTION WITH DOMESTIC WASTE WATER.

Increasing rate of urbanization has lead to increase in fresh water demands for various purposes, domestic, industrial, commercial and others. However only 15-25 % of urban water is consumed and the rest is being returned as wastewater to the hydrologic system. Domestic wastewater containing a large amount of macro nutrients like nitrogen, phosphorous and potassium (NPK) and also a variety of micro nutrients can be used to irrigate (fertigate) forest plantation and trees (Vasudevan et al., 2010). This solves the twin problem of wastewater disposal while supplying valuable nutrients and water to plants and thus enhancing biomass production. Biomass is used not only for renewable energy generation but also as chemical feed stock. Different types of biomass being used for energy generation are woody biomass, grasses and agro residues (McKendry, 2002a, 2002b). A study was undertaken on fertigation of selected trees by domestic wastewater. Results on the growth, pollutant removal, and energy content of trees irrigated by wastewater (W) and ground water (G) are compared.

1.1 Biomass production

A study on the use of grey water through a land treatment system by using trees as vegetation filters was conducted by Indian Institute of Technology, Delhi and G.B. Pant University of Agriculture and Technology (GBPUAT), Pantnagar, Uttarakhand, India at the campus of GBPUAT. The site falls in the sub-humid, subtropical climate zone. The annual average temperature varies from minimum 5.4 ◦C to maximum 41.9 ◦C with average annual rainfall is around 1350 mm.

Here short rotation fast growing tree species were grown in an experimental field using grey water (wastewater) from residential colonies of the University Campus for irrigation. Details on field lay out, spacing etc. are described in an earlier paper ((Pandey et al., 2011). Ground water drawn from a bore well nearby was used as control. Four plant species, eucalyptus (*Eucalyptus hybrid*), poplar (*Populus deltoides*), salix (*Salix alba*) and melia (*Melia azedarach*) were grown. Serpentine, channel and flood irrigation methods were adopted as shown schematically in figure 1. For each species, biomass samples were harvested for analysis from each of the plots of trees grown using grey/wastewater (W) and ground water (G). Results of only the serpentine irrigated samples are presented herein.



f

FIGURE 1. Field lay out for the three modes of irrigation: serpentine, channel and flood.

1.2 Biomass characterization

The samples were shade dried, chipped finely in a rotational lathe machine and pre-dried at 60°C for 72 hrs. They were then milled and sieved through 30 mesh sieve for further investigations. The powdered biomass of all samples was analyzed for moisture and ash content using the standard procedures described in ASTM E871-82 (2006) and ASTM E1755-01 (2007) respectively. Elemental analysis was carried for ultimate analysis. The presence of common elements such as C, H and N, in biomass samples were determined by Elementar vario EL III CHN analyzer. O (wt %) was calculated as the difference of C, H, N and ash from 100. Major inorganic elements in biomass samples were resolved using the acid digestion method (HNO_3) for multi-elemental analysis through a Perkin Elmer Optima 8000 ICP-OES (Inductively Coupled Plasma-Optical Emission Spectrometry).The calorific values for biomass were determined experimentally in a static bomb calorimeter using the procedures described by Anon et al., 1995. Higher heating value (HHV) of samples was theoretically calculated using the modified Dulong's formula (Theegala and Midgett, 2012) as mentioned below.

$$HHV(MJ/kg) = \frac{33.5 \times wt\%\ C}{100} + \frac{142.3 \times wt\%\ H}{100} - \frac{15.3 \times wt\%\ O}{100}$$

Fiber content was determined gravimetrically where finely ground material was analyzed for neutral detergent fiber (NDF), acid detergent fibre (ADF) and lignin following Van Soest method (Burnitt et al., 1985, Guidi et al., 2009). Hemi cellulose was calculated as NDF minus ADF and cellulose as ADF minus lignin.

The ATR/FTIR spectra of biomass samples were obtained to characterize the organic functional groups present in the samples. Perkin Elmer Spectrum One model using the attenuated total reflectance (ATR), was used and values are expressed as γ_{max} cm^{-1}.

1.3 Results and Discussion

For all the plant species the results of proximate, ultimate analysis, calorific value estimation, heavy metal analysis and fiber content are given in table -1, 2 and 3 respectively. As seen from table-1, moisture content of all the samples varied between 6.8 to 7.8% and ash content varied between 1.5 to 3.8 %. Combustibles which included volatile matter as well as fixed carbon varied between 96.1 to 98.4%.

The experimental and theoretical calorific values agree within 1% for all the samples. The biomass grown using wastewater had significantly more calorific value than biomass grown in ground water as control. The experimental calorific value of plants using wastewater for their growth was highest for *E. hybrid* (19.7MJ/kg), and 18.9MJ/kg for *M.azedarach* and *S.alba*, and 17.9 MJ/kg and for *P. deltoids.* The calorific value of plants grown using control water ranged between 17.4-15.6MJ/kg. Percentage increase in calorific value due to waste water irrigation was 6.7% for *P. deltoides*, 8% for *M. azedarach*, 10.4% for *S. alba* and 16% for *E. hybrid*. Plants which have higher calorific value had higher % of carbon content, 50.1% for *E. hybrid*, 49.2% for *M. azedarach,* 49.0% for *S. alba* and 48.2% for *P. deltoides*.

Heavy metal analysis of all the wood samples showed the presence of metals like copper (Cu), lead (Pb), zinc (Zn), iron (Fe), nickel (Ni), chromium (Cr), magnesium (Mg), and manganese (Mn) (Table-2). The metal content in all the plant species was the highest for Fe and Mg followed by Mn, Zn and Cu. The toxic metals like Pb, Ni and Cr were taken up to a small extent 0.5- 3 mg/kg. Cadmium (Cd) and cobalt (Co) were not present in any of the samples or were below detection level of the instrument.

Chemical analysis for fiber content was carried out for all the samples (Table-3). Neutral detergent fiber (NDF) is a measure of hemicelluloses, cellulose and lignin, acid detergent fiber (ADF) is a measure of cellulose and lignin and acid detergent lignin (ADL) is a measure of lignin in the material. Difference of NDF and ADF gives hemicelluloses. %NDF ranged between 91-86, %ADF ranged between 82-68 and %ADL between 26-12. It is seen that lignin content was in the order *M. azedarach* > *P. deltoids* ≥ *E. hybrid* > *S. alba*.

Thermo gravimetric analysis (TGA) in the inert atmosphere of N_2 was carried out to evaluate the pyrolytic degradation and decomposition behavior of the lignocellulosic biomass species. The pattern of weight loss was similar for all the plant species. TGA analysis showed that biochar left was between15-19% in nearly all the samples.

No significant difference could be noted either in thermal degradation or FTIR data between the species and between wastewater (W) and control (G).

1.4. Biomass production and total energy

Studies carried out at the given site by Pandey and Srivastava, 2012, showed that the total dry biomass yield of individual tree in wastewater irrigated plot (treatment) was much more than the yield in control plot. The average increase by fertigation was 2-4 times the dry biomass per tree depending on the plant species. The growth in ground water of the plants was in the order *S. alba*< *M. azedarach* < *E. hybrid* < *P. deltoides*. Biomass production increased by waste water irrigation in the order *M. azedarach* (4.2 times), *S. alba* (3.2 times), *E. hybrid* (2.4 times) and *P. deltoides* (1.4 times) indicating the response of plants to fertigation. Thus plants using wastewater not only had higher biomass yield, but also higher calorific value; therefore the total energy produced is much higher. While higher tree biomass yields in wastewater irrigation have been reported in literature, reports on the increase in calorific value are scarce.

TABLE 1 Proximate and ultimate analysis of biomass irrigated by wastewater (W) and ground water (G) as control

Abbreviation: MW, *M. azedarach* wastewater; MG, *M. azedarach* ground water; PW, *P. deltoids* wastewater; PG, *P. deltoids* ground water; EW, *E. hybrid* wastewater; EG, *E. hybrid* ground water; SW, *S. alba* wastewater; SG, *S. alba* ground water

[a] on dry wt. basis; [b] Including volatile matter and fixed carbon; [c] By difference

Analysis	MW	MG	PW	PG	EW	EG	SW	SG
Proximate analysis (wt %)								
Moisture	7.62	7.86	6.87	7.19	7.75	7.43	6.92	7.39
Ash[a]	1.60	2.23	1.52	2.70	3.88	2.26	2.40	2.18
Combustible[a,b]	98.4	97.77	98.48	97.3	96.12	97.74	97.6	97.82
Ultimate analysis (wt%)[a]								
C	49.16	46.20	48.22	45.68	50.10	46.05	48.99	45.52
H	6.34	6.32	6.26	6.25	7.13	6.21	6.33	6.48
N	0.12	0.15	0.0	0.0	0.15	0.0	0.0	0.0
O[c]	42.78	45.11	43.99	45.37	38.74	46.48	42.28	45.35
Atomic Ratio								
O/C	0.652	0.732	0.684	0.744	0.579	0.757	0.647	0.747
H/C	1.547	1.641	1.557	1.641	1.707	1.618	1.550	1.708
Calorific value (MJ/kg)[a]								
Experimental	18.87	17.43	17.92	16.78	19.67	16.95	18.87	17.08
Theoretical, HHV	18.90	17.52	18.28	17.21	20.96	17.10	18.90	17.64

TABLE 2. Heavy metal analysis of plants irrigated by wastewater

Metals (mg/kg)	MW	PW	EW	SW
Cu	4.76	4.13	20.0	6.398
Pb	1.07	0.91	3.0	1.454
Zn	10.0	6.25	192.2	33.0
Fe	2997	494	166	1408
Ni	0.58	0.50	0.708	0.872
Cr	1.65	1.01	0.53	1.16
Mg	311	542	548	416
Mn	12.9	5.04	12.0	8.8

1.5. Conclusion

Biomass energy output increased by wastewater irrigation as compared to ground water irrigation for all the 4 species of plant namely eucalyptus (*Eucalyptus hybrid*), poplar (*Populus deltoides*), salix (*Salix alba*) and melia (*Melia azedarach*). It was seen that lignin content was melia > poplar > eucalyptus > salix. Carbon content was more in plants grown using wastewater and varied in the order eucalyptus > melia > salix > poplar. The calorific value determined by bomb calorimeter was close to the theoretical value and was higher for all species by 8-10% in wastewater irrigated plants. The increase was in the order eucalyptus > melia > salix > poplar. This study thus indicates that besides biomass yield, the calorific value and hence overall energy output increases by fertigation of fast growing short rotation trees.

Since there was no significant amount of toxic heavy metals in biomass irrigated with wastewater, it can be recommended for irrigation of energy crops for higher biomass yield and energy output, while solving the problem of wastewater disposal and protecting water bodies receiving the wastewater.

TABLE 3. Chemical analysis for fiber content of plants irrigated by wastewater (W) and ground water (G) as control

Plant	MW	MG	PW	PG	EW	EG	SW	SG
%NDF	91.59	89.51	92.54	89.74	85.28	86.17	89.73	87.55
%ADF	75.15	78.47	72.67	68.55	72.45	75.91	71.71	69.67
%Hemi cellulose	16.43	11.04	19.86	21.19	12.83	10.26	18.02	17.88
%Cellulose	53.73	52.26	57.08	51.77	56.28	61.20	57.47	57.36
%Lignin	21.43	26.21	15.60	16.78	16.17	14.71	14.24	12.31

Modified overland flow land treatment system. International Journal of Environmental Sciences 3(1), 591-604.

2. FLOATING WETLAND DESIGN AND INTEGRATION INTO A HYBRID SYSTEM

Advancements have been made in the treatment of water bodies by Floating wetlands which are floating platforms with a bed of media which support deep rooting hydrophytes. Floatation devices such as polystyrene foam blocks, gas-filled tanks and pipes have been tried. Rafts have been made from iron and/or timber frames supported by sealed plastic float tanks or Styrofoam made floating platforms using PVC pipe, chicken wire mesh and fibrous matting material. Low cost rafts have been made of large diameter bamboo interwoven with mats of natural fiber. Plants selected must have the capacity to grow hydroponically and remove various types of contaminants from wastewater within a short contact time and be preferably of economic value. When rafts are floated in wastewater ponds, roots and rhizomes of plants grow into water providing a large surface area for support of bio-films which treat water. The root system also helps by uptake of contaminants and sedimentation of suspended solids. Thus plants play a major role in this treatment. These are being used for treatment of stagnant as well as slow flowing waste water (WW) (Billore et al. 2008, Zhao et al., 2012, Weragoda et al., 2012, Chua et al., 2012, De Stefani et al., 2011, Hamagami et al., 2011, Headley and Tanner, 2006, Wang Chih et al 2013, 2014, Sun et al 2009).

Canna was chosen for the present work because this is a commonly available plant which is compatible with the local agro climates and grows in all seasons in the Indian sub continent. Also the plant does not escape in the environment to become a weed. A number of variants and species of this plant are available and the flowers are aesthetically appealing. The plant is of economic value and is sold as an ornamental plant. Canna indica was taken on rafts for in depth studies for phytoremediation of various contaminants from waste water

In this paper Standalone floating raft units and hybrid systems integrated with constructed wetlands are examined for domestic waste water treatment.

2.1 Materials and Methods

Floating raft design: A photo of a floating raft designed and fabricated is shown in Plate1. The square frame is constructed by joining 4 PVC tubes of length 152 cm and diameter 11 cm. Nylon wire mesh is fixed to the frame for inserting and supporting plants. A set of 4 smaller PVC tubes closed at ends were crisscrossed with another set of 4 tubes. These smaller tubes were of length 174 cm and diameter 2 cm. The trellis work formed thus will help in evenly distributing the load of the plants on the frame. On this frame about 80 plants of canna are taken on coir base, at a spacing of 10 cm between two saplings. The Buoyancy corresponding to this is approximately 62 kg. It may be noted that the raft weight is about 15 Kg. After testing this design, for further work several units of required length and breadth were made and populated by selected plants.

2.2 Site and setting up the field units (Construction of floating raft tank I & II and constructed wetland)

Construction of main *floating raft tank-I & II and constructed wetland were done as shown in the sketch. (Fig.2) The size of floating raft tank-I & II are 14 x 2 x 1.5 (*length *x* width *x* depth*) meters and the size of constructed wetland was kept 10 x 5 x 0.75 meters. Constructed wetland* (CW) was properly lined with high density polythene (HDPE) to prevent the percolation/infiltration of domestic wastewater toward the underground. The relative size width of *floating raft tanks and constructed wetland* were so chosen as to ensure plug flow conditions in the units. The floating wetlands were populated by a variety of plant species. In this paper in only the results pertaining to canna and water lilly floats are presented for pollutant removal in 24 and 48 hour.

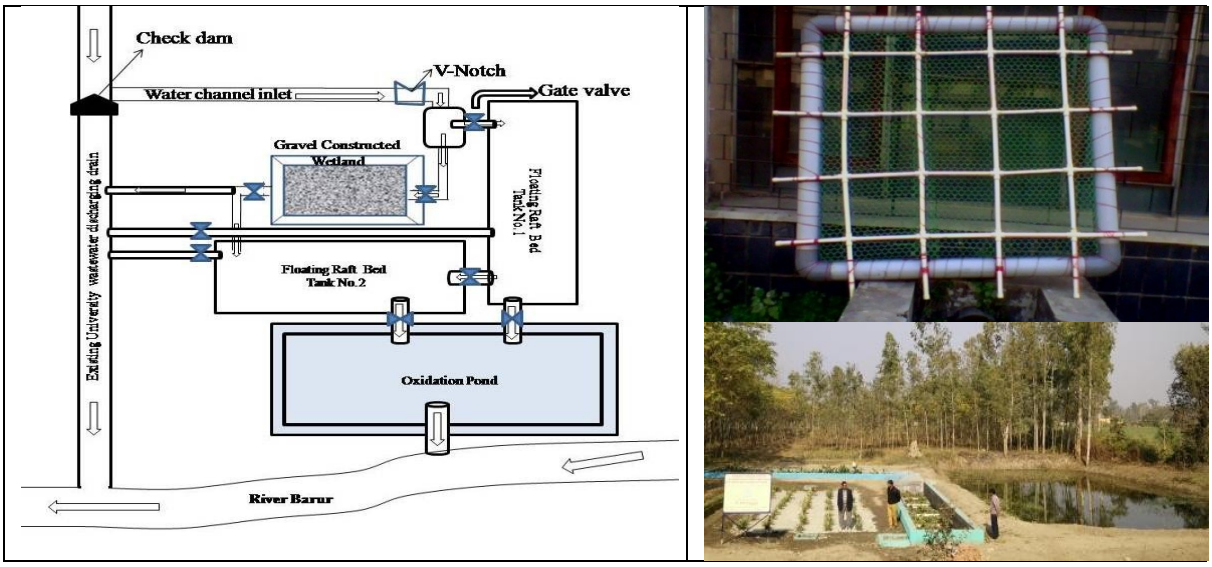


FIGURE 2. Left and right bottom : View of constructed wetland, floating raft tanks and oxidation pond at Pantnagar site; Top right: Frame of floating raft

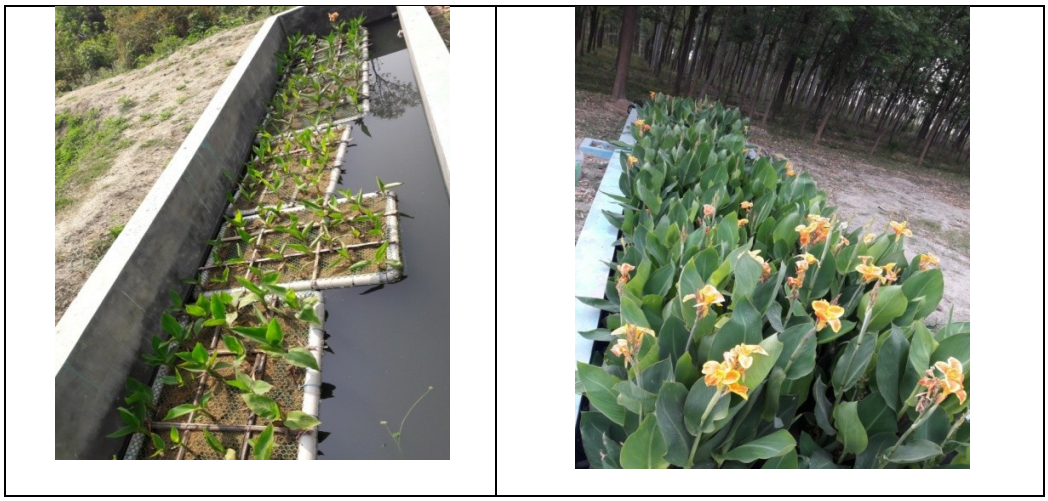


FIGURE 3. Canna (FRT-I) after 5 days of plantation (left) and after 60 days of transplantation (right)

TABLE 4. Pollutants removal in floating raft tank- I (FRT-I) planted with Canna at different HRTs

Parameters	6 Hours HRT		12 Hours HRT		24 Hours HRT		48 Hours HRT	
	Initial	Final	Initial	Final	Initial	Final	Initial	Final
pH	7.5	7.4	7.9	7.6	7.2	7.2	7.7	7.3
EC ($\mu S\ cm^{-1}$)	680	620	710	640	590	549	585	520
K (ppm)	12	9	10	5	11	5	11	4
Na (ppm)	32	23	28	9	31	11	22	6
TS (ppm)	637	492	790	491	765	433	815	409
TDS (ppm)	425	332	530	328	510	291	545	273
TSS (ppm)	212	160	260	163	255	142	270	136
BOD (ppm)	71	56	68	44	72	35	66	27
COD (ppm)	280	210	301	198	280	170	290	150
PO_4^- (ppm)	12.06	10.26	9.81	7.84	11.67	8.64	10.73	5.93
NH_4^+ (ppm)	14.86	11.66	15.30	10.81	13.87	8.27	12.93	6.53
NO_3^- (ppm)	14.19	11.07	13.29	9.84	14.88	9.32	13.34	7.34
TKN (ppm)	24.83	19.87	22.86	14.85	25.19	14.36	23.35	10.19

FRT = Floating raft tank, HRT = Hydraulic retention time

2.3 Wastewater samples collection and analysis.

Domestic wastewater samples collected from study site were analyzed for physico-chemical parameters such as pH, Electrical Conductivity (EC), Chemical Oxygen Demand (COD), Biochemical Oxygen Demand (BOD), Total Solids (TS), Total Dissolved Solids (TDS), Total Suspended Solids (TSS), Nitrate (NO3-), Ammonical Nitrogen (NH4+), Phosphate (PO4-), Potassium (K+), Sodium (Na) and Total Kjeldahl Nitrogen (TKN). For sample collection, preservation and analysis, general guidelines given in Standard Methods for the Examination of Water and Wastewater (APHA, 2012), was followed.

The data on changes in the physicochemical parameters of wastewater on passage through various tanks at different retention times are presented in the tables Pollutants removal in floating raft tank- I (FRT-I) planted with Canna at different HRTs are shown in table 1

In the Hybrid system-I, the two floating raft tanks (I & II) were considered together, where wastewater was passed firstly to the floating raft tank-I planted with Canna and after that it was passed through the floating raft tank-II planted with Water lily. Water samples were collected after passing water through FRT-II and analyzed for water pollutant reduction at different HRT which are shown in Table 5.

TABLE 5. Pollutants removal by FRT-I (Canna) and FRT-II (Water lily) at different HRTs

Parameters	6 Hours HRT		12 Hours HRT		24 Hours HRT		48 Hours HRT	
	Initial	Final	Initial	Final	Initial	Final	Initial	Final
pH	7.5	7.2	7.1	7.1	7.9	7.7	7.3	7.1
EC ($\mu S\ cm^{-1}$)	617	581	593	509	610	570	577	490
K (ppm)	12	8	7	4	8	4	12	3
Na (ppm)	22	14	21	7	25	6	28	6
TS (ppm)	803	515	731	337	737	238	768	262
TDS (ppm)	531	346	482	217	492	163	512	180
TSS (ppm)	272	169	249	120	245	75	256	82
BOD (ppm)	71	53	69	33	60	19	68	14
COD (ppm)	290	170	310	180	260	110	280	80
PO_4^- (ppm)	8.44	6.34	11.20	5.60	7.99	2.16	9.54	2.10
NH_4^+ (ppm)	14.74	9.94	13.03	5.53	12.96	3.96	14.43	3.18
NO_3^- (ppm)	13.03	9.39	14.84	6.68	12.83	4.57	11.98	3.36
TKN (ppm)	24.89	17.97	23.95	10.15	22.81	7.10	25.21	4.81

FRT = Floating raft tank, HRT = Hydraulic retention time

2.4 Biofilms formation on the surface of submerged plant parts

In contrast to the soil system, in wastewater treatment by floating raft system plant gets enough place to develop their root hair zone beneath the raft and shown in Figure 6. The submerged root, stem and leaves in water system beneath the raft are believed to play a crucial role in degradation of many pollutants, since biofilm on the root surface provide the enough surface area for development of different kind of microorganisms which are leads to break down of pollutants into simple form. The root system beneath the rafts were effective at physically trapping of various particulates within the water column, which subsequently slough off the roots as heavy particles that settle down. The enhanced biofilm processes afforded by FRWs may make them particularly effective at removing these pollutants.

A B

FIGURE 4. Developed root surface biofilms beneath the flooting raft a) watger lily b) canna

Similar results have been reported by one of the authors (Vasudevan etal) using small floating rafts made of thermocol populated by canna.

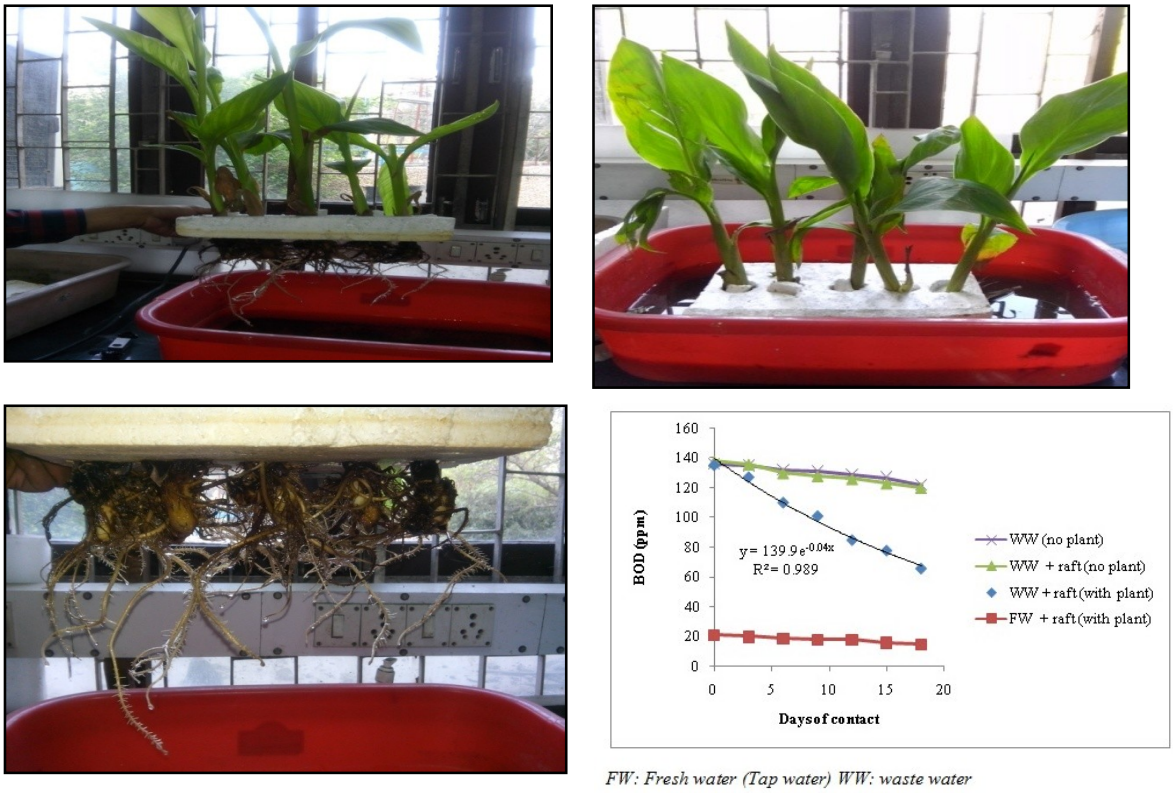


FIGURE 5. Experiments on Canna on thermocol floating rafts

As can be seen in fig. 1 and tables 1 and 2 only a small reduction in BOD was seen in waste water alone and raft without plant in waste water. This reduction may be attributed to exposure to air in the atmosphere. This shows that the raft material does not have significant influence on the BOD. Possibly, there was not enough time for bacterial film formation on the raft surface. Also no mat material was introduced in the raft as the aim was to study the effect of plants alone. In tap water (FW) very little BOD, COD change was seen even with plant (table 2). In comparison there is a significant change in BOD and COD of waste water with plants indicating their important role in phytoremediation. The rate of reduction follows a first order equation.

2.5 Conclusion

Canna rafts are good for domestic waste water treatment by phytoremediation. With development of extensive root system, and increase in root density 50% reduction could be achieved within 2-3 days of contact. The kinetics of removal of BOD is seen to be nearly first order. By immobilizing suitable bacteria on the raft it is possible to achieve high rates of removal. For example, bacterial consortia developed for rapid composting would help.. Further the system is useful for pathogen reduction. It may be noted that the floating raft can also be used to immobilize bio pesticides. For example, immobilizing *Bacillus thuringiensis* would lead to mosquito control. The above study will be useful in designing floating raft system for a village pond in a given locale with an estimation of the amount of waste water accumulating and its characteristics. It is expedient to work with hybrid systems. For example floating raft system can be preceded by a constructed terrestrial wetland.

ACKNOWLEDGMENT

The project support of EPSRC, RCUK , and DST India and RGDWM, India is gratefully acknowledged. The authors from the Indian Institue of Technology Delhi, GBPUAT, Pantnagar, Delhi Technological University, and Aston University U.K. thank these institutions for support.

REFERENCES

1. Anon, J.A.R., Lopez, F.F., Castineiras, J.P., Ledo, J.P., Regueira, L.N., 1995. Calorific values and flammability for forest wastes during the seasons of the year. Bioresource Technology 52, 269–274. doi:10.1016/ 0960-8524(95)00034-C
2. ASTM E871-82 (2006). Standard test method for moisture analysis of particulate wood fuels. ASTM International, Pennsylvania.doi:10.1520/E0871-82R06
3. ASTM E1755-01 (2007). Standard test method for ash in biomass. ASTM International, Pennsylvania. doi:10.1520/E1755-01R07
4. Burritt, E.A., Bittner, A.S., Street, J.C., Anderson, M.J., 1985. Comparison of laboratory methods for the prediction of in vitro dry matter digestibility in three maturing grasses. Journal of Agricultural and Food Chemistry 33(4), 725-728.
5. Guidi, W., Tozzini, C., Bonari, E., 2009. Estimation of chemical traits in poplar short-rotation coppice at stand level. Biomass and Bioenergy 33(12), 1703-1709.
6. McKendry, P., 2002a. Energy production from biomass (part 1): overview of biomass. Bioresource Technology 83(1), 37-46.
7. McKendry, P., 2002b. Energy production from biomass (part 2): conversion technologies. Bioresource Technology 83(1), 47-54.
8. Pandey, A., Singh, M., Srivastava, R.K., Vasudevan, P., 2011. Pollutant removal potential growth and nutritional characteristics of short rotation woody crops in grey water vegetation filter system. J. Sci. Ind. Res. 70, 610-615.
9. Pandey, A., Srivastava, R.K., 2012. Wastewater treatment efficiency and biomass growth of short rotation bio-energy trees in modified overland flow land treatment system. International Journal of Environmental Sciences 3(1), 591-604.
10. Theegala, C.S., Midgett, J.S., 2012. Hydrothermal liquefaction of separated dairy manure for production of bio-oils with simultaneous waste treatment. Bioresource Technology 107, 456–463. doi:10.1016/j.biortech.2011.12.061
11. Vasudevan, P., Thapliyal, A., Srivastava, R.K., Pandey, A., Dastidar, M.G., Davies, P., 2010. Fertigation potential of domestic wastewater for tree plantations. J Sci Ind Res 69, 146-150
12. Billore, S.K., and Sharma, P., 2006. Floating island for treatment of polluted water. Appropriate Technologies 34 (1) 53-54.
13. Billore, S.K., Prashant, and Sharma, J.K., 2008. Restoration and conservation of stagnant water bodies by gravel-bed treatment wetlands and artificial floating reed beds in tropical India, In Proceeding Taal 2007: The 12th World Lake Conf, (M.Sen gupta & R. Dalwani, ed.) 981-987.
14. Chua, H.C.L., Tan, S. B. K., Sim, C.H., Goyal, M. K., (2012) Treatment of baseflow from an urban catchment by a floating wetland system.
15. De Stefani, G., Tocchetto, D., Salvato, M., Borin, M., 2011. Performance of a floating treatment wetland for in-stream water amelioration in Ne Italy. Hydrobiologia 674 (1), 157–167.
16. Hamagami K., Fujihara, M., Edward, L., 2011. Nutrient Removal Rate and Heat Flux through Floating Canna Culture System in Eutrophied Water, Journal of Water and Environment Technology 9(3), 311-320.
17. Headley, T.R., Tanner, C.C., 2006. Application of Floating Wetlands for Enhanced Storm water Treatment: A Review. Auckland Regional Council, Auckland (ARC Technical Publication No. 324).
18. Sun, Lianpeng, Liu, Yang, and Jin, Hui, 2009. Nitrogen removal from polluted river by enhanced floating bed grown canna. Ecological Engineering, 35, 135-140.
19. Wang, Chih, Yu, Sample, David J., 2013. Assessing floating treatment wetlands nutrient removal performance through a first order kinetics model and statistical inference. Ecological Engineering, 61, 292-302.

20. Wang, Chih, Yu, Sample, David J., 2014. Assessment of the nutrient removal effectiveness of floating treatment wetlands applied to urban retention ponds. Journal of Environmental Management, 137, 23-35.
21. Wang, Chih, Yu, Sample, David J., Day, Susan D. And Grizzard, Thomas J., 2014. Floating treatment wetland nutrient removal through vegetation harvest and observations from a field study. Ecological Engineering (in print).
22. Weragoda, S. K. Jinadasa, K. B. S. N., Zhang, D. Q., Gersberg, R. M., Tan, S. K., Tanaka, N., Jern, N. W., (2013). Tropical Application of Floating Treatment, Wetlands, DOI 10.1007/s13157-012-0333-5.
23. Zhao, F.L., Xi, S.ng, X.E., Yang, W.D., Li, J.J., Gu, B.H., He, Z.L., 2012. Purifying eutrophic river waters with integrated floating island systems. Ecol. Eng. 40, 53–60
24. Padma Vasudevan Sena*, Monica Vermab, Satyawati Sharmab, P.K. Sena and S.N. Singhaa, A Study on Floating Rafts with Canna Plant for Domesticwater Treatment, .Int. J. Environmental Sciences Vol 3 (4) : pp 221 - 228 (2014)

Materials Research Express

ACCEPTED MANUSCRIPT

Effect of direct LN2 single jet supply during turning of AISI D3 steel alloy using an optimization technique

To cite this article before publication: Anurag Sharma *et al* 2019 *Mater. Res. Express* in press https://doi.org/10.1088/2053-1591/ab2ded

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 06/07/2019 at 05:57

Effect of direct LN2 single jet supply during turning of AISI D3 steel alloy using an optimization technique

Anurag Sharma[1], R.C. Singh[2], Ranganath M.Singari.[2]

[1]Research Scholar, Department of Mechanical Engineering, Delhi Technological University, Delhi-110042.India

[2]Professor, Department of Mechanical Engineering, Delhi Technological University, Delhi-110042.India

Corresponding author: Ranganath M.Singari, Department of Mechanical Engineering, Delhi Technological University, Delhi-110042.India, email id ranganath@dce.ac.in

Keywords: Cryogenic turning, dry turning, optimization, ANOVA, liquid nitrogen

Abstract: The emerging demand for present and future research work to environment conscious machining is high surface integrity of machined workpiece. Researchers are continuously working to meet intrinsic and strict international standards with close tolerance limits for the fabrication of hard engineering materials. Cryogenic cooling with liquid nitrogen at the interface of tool and workpiece may provide a better alternative than conventional cutting fluids during turning operation. Design of experiments using Taguchi based L_{18} in dry and LN2 machining conditions has been used to carry the experiments with TiN coated tungsten carbide cutting inserts on the workpiece of AISI D3 steel alloy. The separate probability plot of surface roughness, cutting force, flank wear length and temperature at 95% of confidence level have been analysed for optimization and investigation of the experiment. ANOVA has been used to determine the effect of the contribution of control factors. The machining condition has the highest effect on the percentage contribution for surface roughness, cutting force, flank wear length and temperature as 62.42%, 59.77%,

Accepted Manuscript

70.03% and 49.69% respectively. The other contributing factors in order of decreasing effect have been observed as cutting speed, feed and depth of cut for machining characteristics. The predicted values at the optimized level have been confirmed by the actual performance of the experimental run. Regression models have been developed. R-Sq values for surface roughness, cutting force, flank wear length and temperature have been found as 83.02%, 84.58%, 77.85% and 81.09%. SEM images of cutting inserts depict the crater wear, built up edge and adhesive wear during dry where as negligible wear in cryogenic atmospheric turning operations. The EDS images of machined workpiece depict the Fe has major composition.

1Introduction:

The prime requirement of modern technology is performing the machining process under low temperature at the interface of tool and workpiece without affecting the natural environment and operator's health. This is a step forward towards Green Manufacturing. Environmentally conscious and energy efficient approach is negligible wastage of tool and workpiece material while achieving highly improved surface finish during machining. This is helpful in the conservation of resources. This is economical in controlling the cost embedded [1-3]. Machining without the use of any coolant (dry machining) is an eco-friendly step but machinability characteristics are not satisfactory at high cutting parameters (speed, feed and depth of cut) [58]. The running of the machine at low cutting parameters is not cost effective due to slow manufacturing rate. The use of conventional cutting fluids makes improvement in machinability characteristics but due development of new hard engineering materials, high close tolerances to be maintained during machining and International Standards of Quality the conventional cutting fluids are not recommended for use [4-8]. The eco-friendly alternative presently available is liquid nitrogen. The boiling point of liquid nitrogen is -196°C. LN2 is colourless, odourless, tasteless and non-toxic. The high heat removal rate

Accepted Manuscript

makes better coolant. The high efficiency and effective heat removing capacity at the interface of tool and workpiece is better for high cutting parameters during machining of hard materials like superalloys, composite material, difficult to machine materials, titanium alloys, steel alloys etc. [9-11]. The chips and debris generated during machining are free from oil, harmful chemicals etc. and can be used as raw material for recycling of scrape. This does not require any additional cost and chemical method in cleaning of chips and debris as compared to conventional cutting fluids. This is cost effective and eco-friendly [12-14]. A neat and clean working environment without any toxic fumes is provided by LN2. This is appreciable good for the operator's health. High chromium tool steel alloy AISI D3 (HRC 60) is used in manufacturing blanking & forming dies, forming tools, press tools, punches, bushes and wear resistant moulds. It is categorised as difficult to machine materials. Cutting parameters for turning of AISI D3 steel alloy is speed 30-260m/minute, feed 0.05-0.15mm/revolution and depth of cut 0.2-0.5mm [15]. MQL (Minimum Quantity Lubrication) is an alternative of conventional water- miscible cutting fluid process. Edible oils like sunflower, canola, olive, castor, coconut oil etc. and non-edible oils like jatropha, kranj, mahua, neem etc. are used. The metered oil (100-200ml/hour) is sprayed at the interface of cutting tool and workpiece. The results of machinability characteristics are good and improved. The lubrication is good but heat removal rate is slow. This is not recommended for high levels of cutting parameters due to a high amount of heat generation [13-14, 21].

1.1 Cryogenic Cooling: Liquid nitrogen is directly supplied in the form of a pressurised jet at the interface of cutting tool and workpiece. Grinding forces obtained during LN2 declined by 32-36% during dry 13-26% during flood grinding and surface roughness declined by 26-59% during dry and 32-43% during flood grinding [37-38]. Cryogenic turning with a direct regular supply of LN2 at flank and rake parts of single point cutting insert and found that cutting temperature declined by 61-66%, cutting force declined by 35-42%, surface roughness and

Accepted Manuscript

tool wear declined by 35% and 39% respectively in flood machining [41]. In a comparative investigation for tool wear and surface roughness during dry, wet, MQL, CO2(rake)+MQL(flank), LN2, LN2(rake)+MQL(flank) found that surface roughness is the lowest with LN2 [43]. A comparative investigation has been made between dry, wet, MQL and LN2 cooling processes. The results obtained during LN2 supply at the interface of cutting insert and workpiece are better [42]. The influence of cryogenic cooling on steel alloys AISI 1040, E4340C, AISI 4140, AISI 4037 has given low tool wear and surface roughness as compared to wet and dry machining. The surface produced during cryogenic machining has been found clear with a very low or negligible amount of debris [44-47]. The machined surface has low cracks. The surface profile graph is uniform with variations of a small range. The chips produced are discontinuous and thin with the clean surface [42, 48 - 51]. S. Bruschi et al. [62] studied the effects of cryogenic turning on AISI 1316L alloy found at lower cutting temperature enhanced compressive residual stress, reduced the formation of defects on machined surface and increment of corrosion resistance. The indirect approach has been used for cooling down the air by LN2 and further chilled air is sprayed at the interface of the cutting tool and surface roughness of workpiece is improved as compared to dry machining. Chip thickness declined by 9 % at cutting speed lower than 50m/minute [36].

The hardness of tool material is increased by cryogenically treatment. This process involves keeping the tool material at -196°C for 12-48 hours according to the type of material. In cryo processors, the cutting tool inserts and cutting tool are placed in a closed chamber. The temperature is gradually decreased from room temperature to -196°C and maintained for 12 hours and gradually increased to room temperature. Grains of material become refined resulting in the increment of hardness and improvement in wear properties. This has been investigated that cryogenically treated cutting inserts coated and uncoated perform better than untreated coated and uncoated cutting inserts respectively during machining operations. Tool

Accepted Manuscript

life and wear resistance have been increased by 20-45% and 85-96% respectively for cryogenically treated cutting inserts [54-57]. Priya Jadhav et al. [63] gave deep cryogenic treatment to coated tungsten carbide inserts for 36 hours and found low tool wear and surface roughness of machined workpiece.

1.2 Optimization:

Machining processes consist of various levels of machining parameters. Optimization saves the economy, resources, raw material, energy and time from the initial experimental stage to industrial mass production. This technique eliminates hit and trial choice of levels in machining processes [26, 30]. Taguchi technique was discovered in 1950 by Genichi Taguchi [16-21]. This is based on (S/N) ratio signal-to-noise. This gives a selection for the best combination of controlled factors (cutting parameters in machining) for the better conditional output of machinability characteristics [33-34,66]. In the present collected literature of research work, Taguchi approach for optimization is one of the classic and reliable technique. In dry turning of hardened steel alloys AISI D3, AISI4340, AISI H13, EN19, EN47 and high-alloy white cast iron the surface roughness, tool wear were optimized by Taguchi approach [22-25, 27-29, 31-32, 60]. Fuat Kara [63] used Taguchi technique with L_9[OA] in optimization of machinability characterstics like surfaceroughness and flank wear during dry turning of DIN 1.2344 hot work tool steel. Fuat Kara et al. [64] investigated the performance of cutting tool material coated with different processes of coating like physical vapour deposition (PVD) and chemical vapour deposition (CVD). Design of experiments was based on L_{18}. It has been found that cutting tool material coated with PVD performed better. Fuat Kara et al. [65] predicted the cutting temperature in orthogonal machining of AISI 316L using artificial neural network.In minimum quantity lubrication technique (MQL) of machining condition with a controlled amount of supply of castor oil gave better results in surface roughness and tool wear [21]. In machining condition of cryogenic with LN2 surface

Accepted Manuscript

roughness, cutting force, coefficient of friction and tool wear are low as compared with dry and conventional cutting fluids [19]. Analysis of variance (ANOVA) is used to study the percentage of contribution effect of machining parameter like machining condition, speed, feed, depth of cut and tool angle in the evaluation of machining characteristics like surface roughness, cutting force, tool wear and cutting temperature [22-24].

Researchers have investigated and presented the results that during machining processes like turning, milling and grinding in machining condition of dry, wet, MQL, cryogenic with LN2 and frozen CO2. The better and improved results are given by cryogenic cooling by LN2. ANOVA models proposed in above mentioned work are efficient between 69-97% at the confidence level of 95%. This work has a future scope of machining with other noble gases like hydrogen, helium and a hybrid combination of gases with MQL. In the presented investigation, turning was done with TiN coated cutting inserts during dry and direct LN2 supply at the interface of cutting insert and workpiece. The results obtained were analysed by Taguchi and ANOVA for machining characteristics.

2 Experimental procedure:

2.1 AISID3 steel alloy in the form of a round bar of a diameter of 40 mm and a length of 250 mm was used as the workpiece. Table 1 shows the percentage of elements by weight.

Table 1 Chemical analysis of $AISID_3$ steel

C	Si	Mn	S	P	Cr	Ni	Mo	Co	Nb	V	W	Fe
2.01	0.249	0.429	0.028	0.021	11.00	0.075	0.09	0.012	0.022	0.037	0.088	85.939

Cutting inserts were used in the shape of diamond of tungsten carbide of thickness 4mm coated with titanium nitride, end relief angle 11°, front rake angle 0°, nose radius 8 mm of

Accepted Manuscript

specification as DCMT11T308HQ grade PV 20 with left hand cutting tool holder of a square face of 12 mm and length 80mm. Table 2 shows the percentage of elements by weight.

Table 2 Chemical analysis of cutting inserts

C	Co	Mn	V	Nb	Ni	W	Mo	Ti
7.09	16.27	0.29	2.06	5.24	4.94	15.89	0.68	47.54

2.2 Machining methods and set up

The experiments were performed on three jaws Harrison centre lathe machine with maximum swing limit of diameter 400mm. Liquid nitrogen dewar container of 55 kg capacity was for the direct supply at the interface of cutting insert and workpiece. After the completion of each experiment, the used cutting insert was placed safely for further measurement. The new unused cutting insert was used for each complete stage of the experiment. Surface roughness was measured by Taylor Hobson Surtronic 3 + surface roughness tester. The surface roughness was measured at five different places on the workpiece for calculating a final average value. The cutting force was measured by a piezo-electric dynamometer. The flank was measured by an optical inspection machine at three different places on a flank and the average value was calculated. The temperature was measured by IR thermal imaging thermometer focused at the interface of rake face of cutting tool insert and workpiece during the end of an experiment.

2.3 Design of Experiment (DoE)

Accepted Manuscript

Machining experiments were performed with four control factors. One factor was varied at two levels and three factors were varied at three levels. Former factor machining condition was varied to two levels (level 1 = dry) and (level 2 = LN2). Cutting parameters three levels of speed (30, 45 and 60 m/min.), feed (0.05, 0.08 and 0.14mm/rev.) and depth of cut (0.25, 0.35 and 0.45). Table 3 shows control factors with different level values. L_{18} orthogonal array [OA] based on Taguchi as shown in Table 4 was for the performance of the experiment. This has reduced the experiments to 50%. This saves energy, the raw material for the workpiece, cutting tool and operator cost.

3 Experimental results and discussion

3.1 Probability analysis

The experiments were performed according to L_{18} Taguchi based design of the experiment. The responses were recorded and listed in table 5. The responses have been justified by the probability plot from Fig. 1(b). The data values are roughly aligned with the middle line and normally distributed. This can be further used for optimization and investigation.

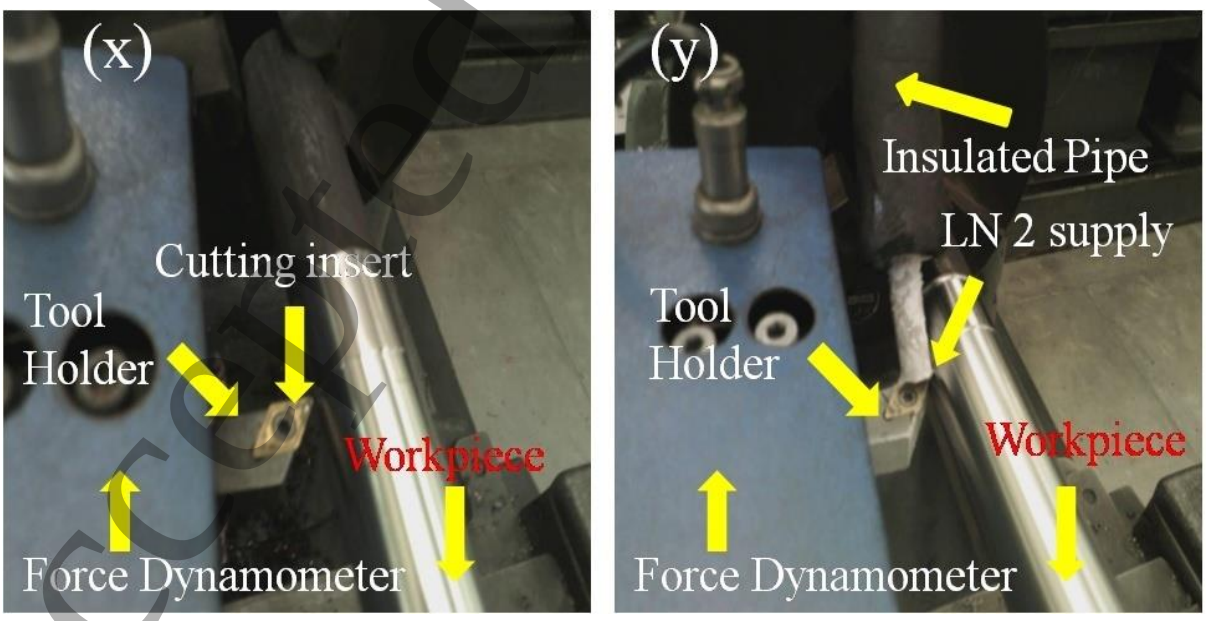


Fig. 1(a) Dry turning{x} and Liquid turning{y} as per L_{18} DoE [OA]

Accepted Manuscript

Nomenclature

Surface roughness, Ra'

Cutting force, Fc'

Flank wear length, Vb'

Temperature, T'

Machining condition, M'

Speed, Vc'

Feed, Fo'

Depth of cut, Dc'

Liquid nitrogen, LN2

Design of experiments, DoE

Accepted Manuscript

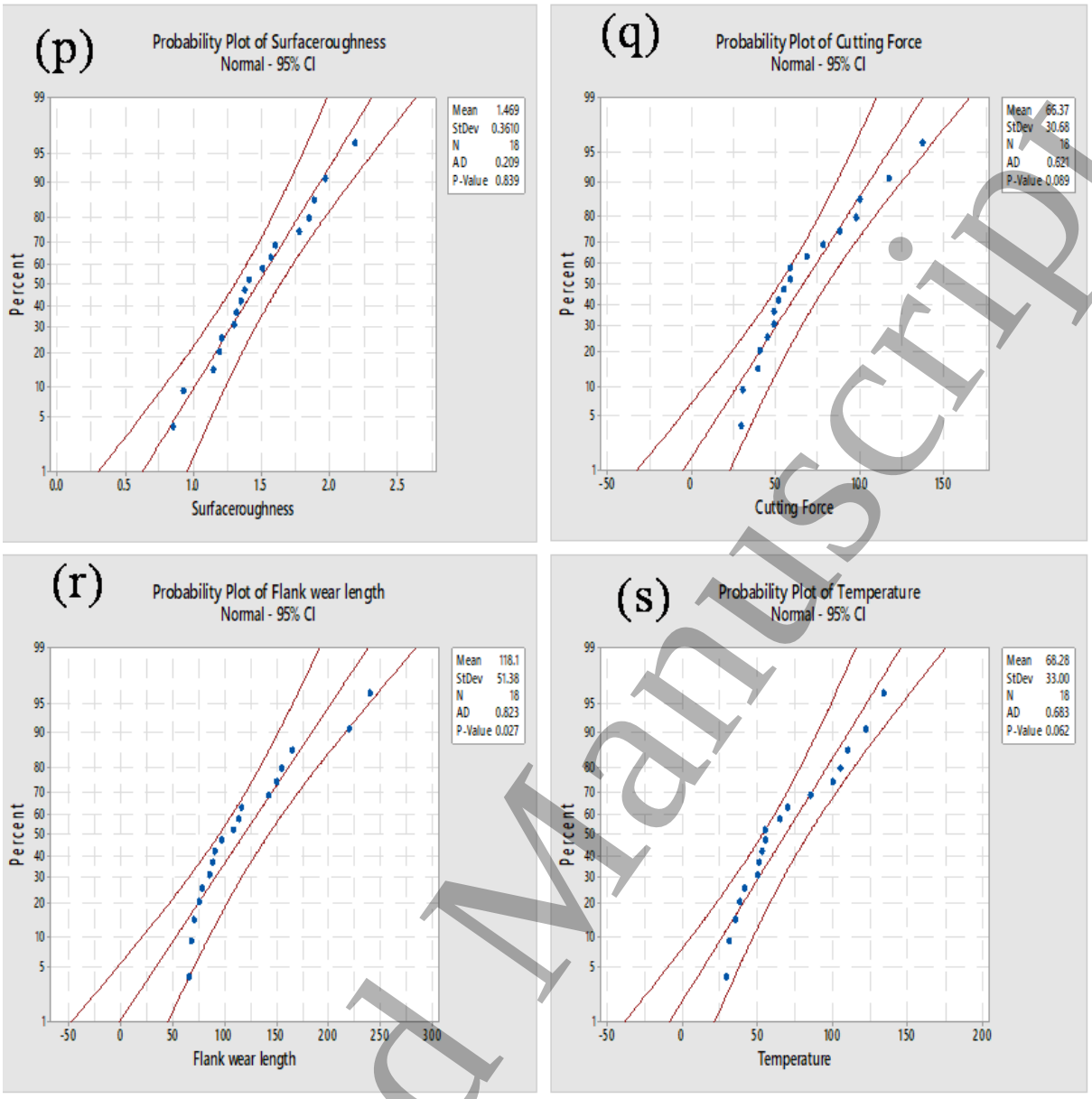


Fig.1 (b) Probability plots for (p) surface roughness, (q) cutting force, (r) flank wear length and (s) temperature

3.2 Taguchi Based Optimization (Signal to Noise ratio)

The optimization based on Taguchi S/N (ratio) involves a reduction of variability and alignment of mean value to target value [21, 27]. In any process, variability may arise due to factors having no control are termed as uncontrollable factor or noise. Eq. (1) shows smaller

Accepted Manuscript

the better characteristic of continuous response function depending on the optimization principle [20, 28].

Smaller is the better characteristic $\frac{S}{N} = -10log\frac{1}{n}(\sum x^2)$ (1)

Where x represents the measuring responses (Ra', Fc', Vb' and T') & n is the number of experimental data. Table 4 shows coded and un-coded values of factors according to L_{18} orthogonal array [OA]. The responses with calculated S/N values are shown in table 5. Taguchi based S/N values of each response are shown in table 6. The difference between the maximum and minimum S/N value of each control factor is shown as delta. The optimised value of each control factor has been selected by recognizing the highest S/N value of each control factor for a particular response of surface roughness (Ra'), cutting force (Fc'), flank wear length (Vb') and temperature (T'). According to the importance of the control factor, Rank value is present. This is supported by the statistical analysis of variance (ANOVA). The table of ANOVA has a degree of freedom (df), adjoint sum of squares (Adj SS), adjoint mean of square (Adj MS), F-Value, P-Value and percentage of contribution. The P-value is defined for the significance level of 5% (confidence level of 95%) for all responses [18, 21]

Table 3 Control factors with different level values

Levels	Machining Condition	Cutting parameters		
		Vc' (m/min.)	Fo' (mm/rev.)	Dc' (mm)
Level 1	Dry	30	0.05	0.25
Level 2	LN2	45	0.08	0.35
Level 3	-	60	0.14	0.45

Accepted Manuscript

Table 4 L_{18} Taguchi orthogonal array

Run	Coded Values of factors				Uncoded values of factors			
	M'	Vc'	Fo'	Dc'	M'	Vc' (m/min.)	Fo' (mm/rev.)	Dc' (mm)
1	1	1	1	1	Dry	30	0.05	0.25
2	1	1	2	2	Dry	30	0.08	0.35
3	1	1	3	3	Dry	30	0.14	0.45
4	1	2	1	1	Dry	45	0.05	0.25
5	1	2	2	2	Dry	45	0.08	0.35
6	1	2	3	3	Dry	45	0.14	0.45
7	1	3	1	2	Dry	60	0.05	0.35
8	1	3	2	3	Dry	60	0.08	0.45
9	1	3	3	1	Dry	60	0.14	0.25
10	2	1	1	3	LN2	30	0.05	0.45
11	2	1	2	1	LN2	30	0.08	0.25
12	2	1	3	2	LN2	30	0.14	0.35
13	2	2	1	2	LN2	45	0.05	0.35
14	2	2	2	3	LN2	45	0.08	0.45
15	2	2	3	1	LN2	45	0.14	0.25
16	2	3	1	3	LN2	60	0.05	0.45
17	2	3	2	1	LN2	60	0.08	0.25
18	2	3	3	2	LN2	60	0.14	0.35

Table 5 Experimental results of responses as per Taguchi L_{18} orthogonal array

Accepted Manuscript

Run	Responses				S/N value of responses			
	Ra'	Fc'	Vb'	T'	Ra'	Fc'	Vb'	T'
1	2.19	59.0	108	38	-6.808	-35.417	-40.668	-31.596
2	1.89	49.0	113	55	-5.529	-33.803	-41.062	-34.810
3	1.97	68.6	116	70	-5.889	-36.726	-41.289	-36.902
4	1.51	78.4	155	85	-3.579	-37.886	-43.806	-38.588
5	1.60	88.2	165	105	-4.082	-38.909	-43.349	-40.423
6	1.85	98.0	150	122	-5.343	-39.825	-43.522	-41.727
7	1.57	117.6	142	134	-3.918	-41.408	-43.046	-42.542
8	1.38	137.2	220	100	-2.796	-42.747	-46.848	-40.000
9	1.78	100.0	240	110	-5.008	-40.000	-47.604	-40.828
10	1.35	29.4	70	29	-2.606	- 29.367	-36.902	-29.247
11	1.41	30.2	85	35	-2.984	-29.600	-38.588	-30.881
12	1.30	39.2	90	41	-2.278	-31.866	-39.085	-32.255
13	0.93	40.5	68	31	0.630	-32.149	-36.65	-29.827
14	1.21	49.0	78	50	-1.655	-33.803	-37.842	-33.979
15	1.32	58.8	65	55	-2.411	-35.388	-36.258	-34.807
16	1.19	51.5	75	65	-1.511	-34.236	-37.501	-36.258
17	0.85	55.0	88	51	1.412	-34.807	-38.889	-34.151
18	1.15	45.0	97	53	-1.214	-33.064	-39.735	-34.485

Accepted Manuscript

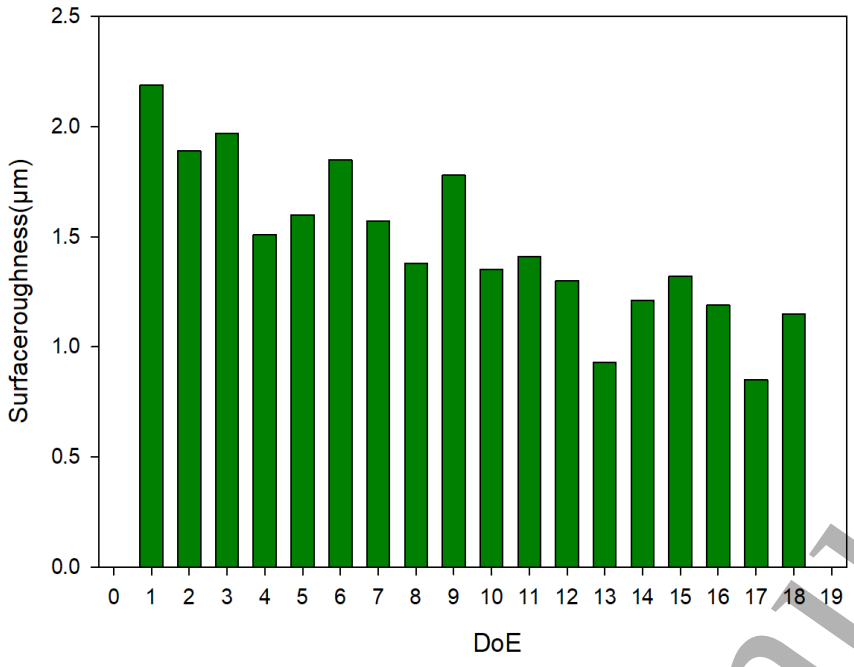


Fig. 2 Experimental Ra' in accordance with DoE

3.3 Surface roughness: Surface roughness is the measure of service period provided by a machined component. Low surface roughness reduces the stress points and makes an increment in the good service life period. This declined the wastage of energy and cost during machining by a high integrity surface of the machined component [18]. Surface roughness during cryogenic turning with a direct supply of LN2 at the interface of cutting tool and workpiece is low due to less adhesion between cutting tool insert and machined surface [41, 43]. Fig. 2 shows the experimental values of surface roughness in different machining conditions (dry and cryogenic with LN2). It has been observed that values of surface roughess in dry machining condition L_{18} DoE (1-9) is more than cryogenic (10-18). This may be due to high friction and heat generated at the interface of tool and workpiece. Fig. 3 (a)

Accepted Manuscript

and Table 6 show the modification of mean s/n ratio of Ra'. Fig. 3 (b) and Table 7 show modification of means Ra'.

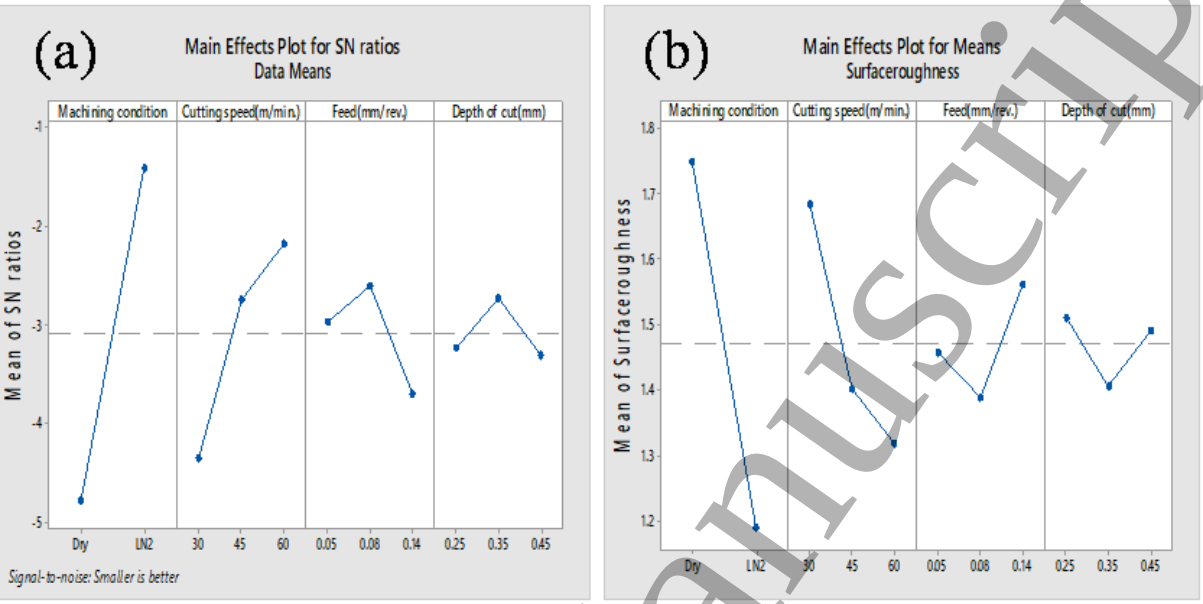


Fig. 3(a) Modification of mean s/n ratio Ra' (b) Modification of mean Ra' with various factor levelsTurning with a direct supply of LN2 at the interface of tool and workpiece declined surface roughness as compared to dry turning. Fig.3(b) shows on incrementing cutting speed surface roughness declined. On incrementing feed and depth of cut surface roughness declined and then increased. The graph shows the representation of both machining condition. Cryogenic cooling with direct supply of LN2 has significant effect on values obtained of surfaceroughness. Table 4, shows optimum level value of surface roughness at machining condition (level 2 = LN2), Vc' (level 3 = 60m/min.), Fo' (level 2 = 0.08mm/rev.) and Dc' (level 2 = 0.35mm). The predicted value of the response is calculated from Eqs. (2) and (3). dp is the S/N ratio calculated at optimum level, $\overline{dp}$ is the average S/N ratios of all variables at optimum level, $\overline{Mo}$ is the average s/n ratio when variable M (machining condition) is at optimum level, $\overline{No}$ is the average s/n ratio when variable N (speed) is at optimum level, $\overline{Fo}$ is the average s/n ratio when variable F (feed) is at optimum level and $\overline{Do}$

Accepted Manuscript

is the average s/n ratio when variable D (depth of cut) is at optimum level, Rp is predicted responses for surface roughness, cutting force, cutting time, flank wear length and temperature. The predicted value of surface roughness has been calculated from Eqs. (2) and (3) is 0.75µm. ANOVA table 8 for response of surface roughness shows machining condition has the highest effect in percentage of contribution (62.42%), speed has a second higher effect on the percentage of contribution (18.67%), feed has a third higher effect on the percentage of contribution (4.47%) and the depth of cut has fourth higher effect on the percentage of contribution (1.40%). F- value shows the relative importance firstly machining condition, secondly speed, thirdly feed and lastly, is depth of cut. P- value is significant for speed. Since P- value is significant at the value of level equal or less than 0.05.

$$dp = \overline{dp} + \left(\overline{Mo} - \overline{dp}\right) + \left(\overline{No} - \overline{dp}\right) + \left(\overline{Fo} - \overline{dp}\right) + \left(\overline{Do} - \overline{dp}\right) \quad (2)$$

$$Rp = 10^{-dp/20} \quad (3)$$

3.4 Cutting Force: Cutting force is created during the machining process is related to machining condition. In dry machining (without any coolant) the magnitude of cutting force is large. In wet machining, the magnitude of cutting force is comparatively low. This may be due to low friction and better cooling and lubrication. But, with the emergence of hard engineering materials, the conventional cutting fluid is not sufficient. The effective alternative is presently liquid nitrogen [39].

Accepted Manuscript

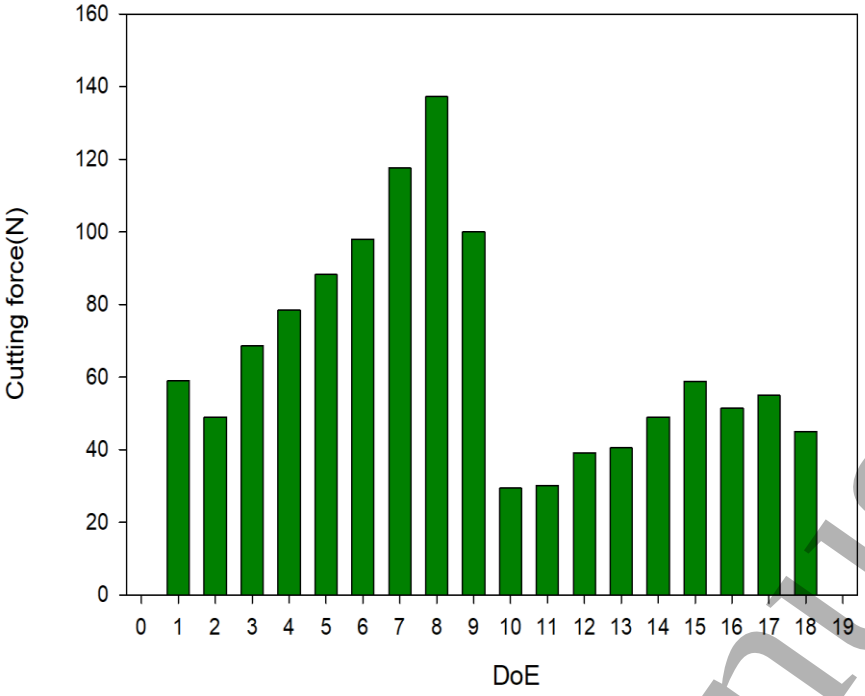


Fig. 4 Experimental Fc' in accordance with DoE

Fig. 4 shows the experimental values of cutting force in different machining conditions (dry and cryogenic with LN2). It has been observed that values of cutting force in dry machining condition L_{18} DoE (1-9) is more than cryogenic (10-18). This may be due to high friction and heat generated at the interface of tool and workpiece. Cryogenic cooling with direct supply of LN2 has significant effect on values obtained during experimentation. Cutting force is low during cryogenic turning with a direct supply of LN2 as compared to dry turning due to better lubrication effect created by liquid nitrogen at tool (rake face) chip interface and newly machined surface with the flank face of the cutting tool due to the formation of the fluid. This has reduced adhesion between interacting surfaces [41, 43]. Fig. 5(a) and Table 6 show modification of s/n ratio of Fc' Fig. 5(b) and Table 7 show modification of means of Fc'.

Accepted Manuscript

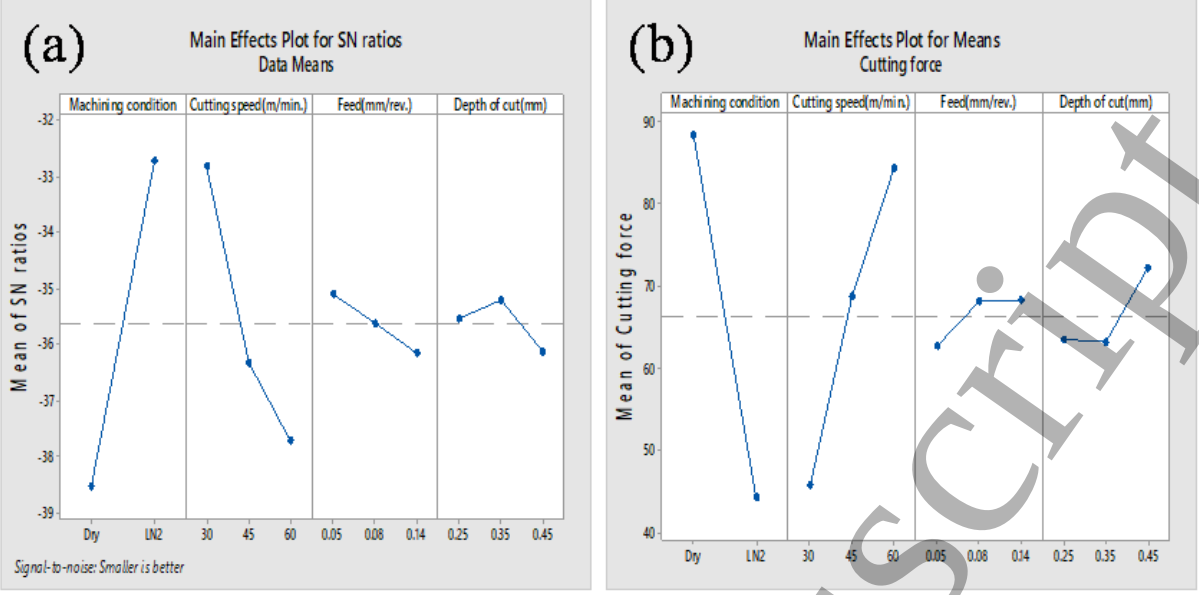


Fig. 5 (a) Modification of s/n Fc' (b) Modification of means Fc' with various factor levels

On incrementing Vc', Fo' and Dc' mean of cutting force has been increased. Table 6 shows optimum level of cutting force at machining (level 2 = LN2), Vc' (level 1 = 30m/min.), Fo' (level 1 = 0.05mm/rev.) and Do' (level 2 = 0.35mm). The predicted value of cutting force has been calculated from Eqs. (2) and (3) is 28.01N. ANOVA table 8 for response of cutting force shows machining condition has the highest effect on percentage of contribution (59.77%), speed has a second higher effect on the percentage of contribution (30.14%), feed has third higher effect on the percentage of contribution (1.34%) and depth of cut has fourth higher effect on the percentage of contribution (1.02%). F- value shows the relative importance firstly machining condition, secondly speed, thirdly feed and lastly is depth of cut. P- value is significant for machining condition and speed.

3.5 Tool Wear: Tool wear is related to machining condition, tool material and workpiece material. Fig. 6 shows the experimental values of flank wear length in different machining conditions (dry and cryogenic with LN2). It has been observed that values of flank wear length in dry machining condition L_{18} DoE (1-9) is more than cryogenic (10-18).

Accepted Manuscript

. This may be due to high the temperature created at the interface of tool and workpiece. Abrasion and adhesion phenomena are present in flank wear mechanism. Flank surface rubbed with the workpiece which acts as a resultant for dimensional change of cutting edge. The mechanism liable for the occurrence of crater wear is abrasion, adhesion and diffusion. The sliding motion of chip over the rake surface generates high temperature favouring diffusion phenomenon [36, 43].

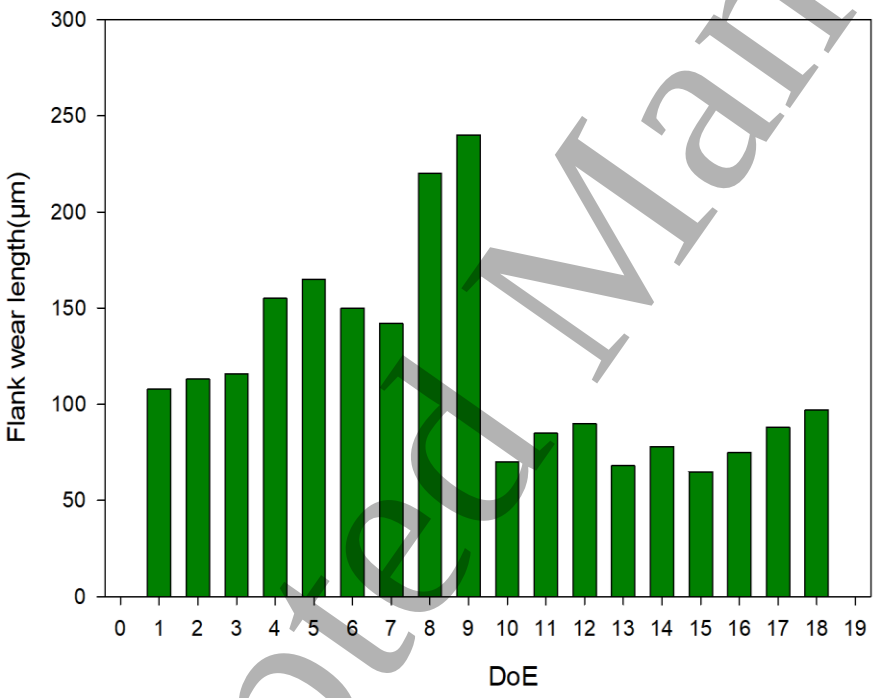


Fig. 6 Experimental Vb' in accordance with DoE

In cryogenic cooling, tool wear is low as compared to dry turning due to low temperature created at the interface of tool and workpiece. This reduced the thermal softening of tool and workpiece material. Less adhesion between interacting surfaces. Fig. 10(a) shows SEM images of the cutting tool insert with large flank wear length during dry turning and Fig. 10

Accepted Manuscript

(b) shows low flank wear length and negligible adhesive during cryogenic turning. Flank wear is created by abrasion and attrition. Flank wear length is the low during cryogenic turning. This may be due to low temperature created and maintained during turning with a direct supply of LN2.

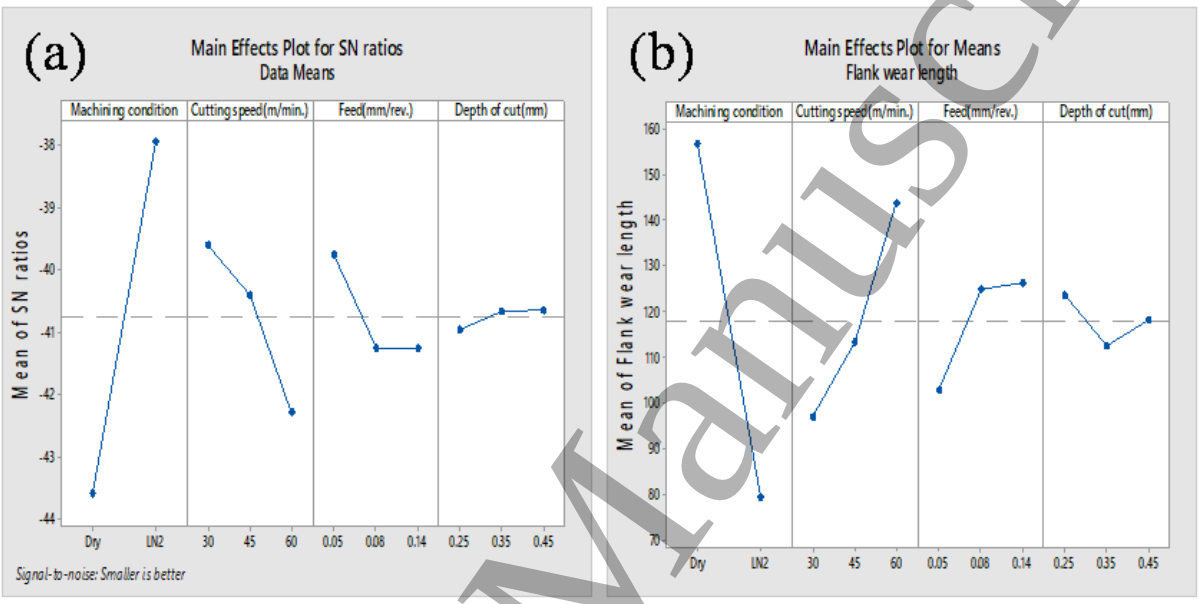


Fig.7(a) Modification of s/n of Vb' (b) Modification of mean Vb' at various factor levelsFig.7(a) and Table 6 show modification of s/n ratio of Vb'. Fig.7(b) and Table 7 show modification of mean of Vb' and on incrementing speed and feed tool wear has increased. On incrementing depth of cut flank wear length marginally decreased than increased. This may be due to a higher cooling efficiency of LN2. At low cutting speed and depth of cut, flank wear is a marginal difference in dry and cryogenic machining. Table 6 shows optimum level of flank wear length at machining (level 2 = LN2), Vc' (level 1 = 30m/min.), Fo' (level 1 = 0.05mm/rev.) and Dc' (level 2 = 0.35mm). The predicted value of flank wear length has been calculated from Eqs. (2) and (3) is 61.97µm. ANOVA table 8 for response of flank wear length shows machining condition has the highest effect on the percentage of contribution (70.03%), speed has a second higher effect on the percentage of contribution (11.89%), feed

Accepted Manuscript

has a third higher effect on the percentage of contribution (4.09%) and depth of cut has fourth higher effect in the percentage of contribution (0.37%) F- value shows the relative importance firstly machining condition, secondly speed, thirdly feed and lastly is depth of cut. P- value is significant for machining condition and speed.

3.6 Temperature: Temperature at the interface of tool and workpiece is important in relating tool wear and surface integrity of the machined component. Fig. 8 shows the experimental values of temperature in different machining conditions (dry and cryogenic with LN2). It has been observed that values of temperature in dry machining condition L_{18} DoE (1-9) is more than cryogenic (10-18).

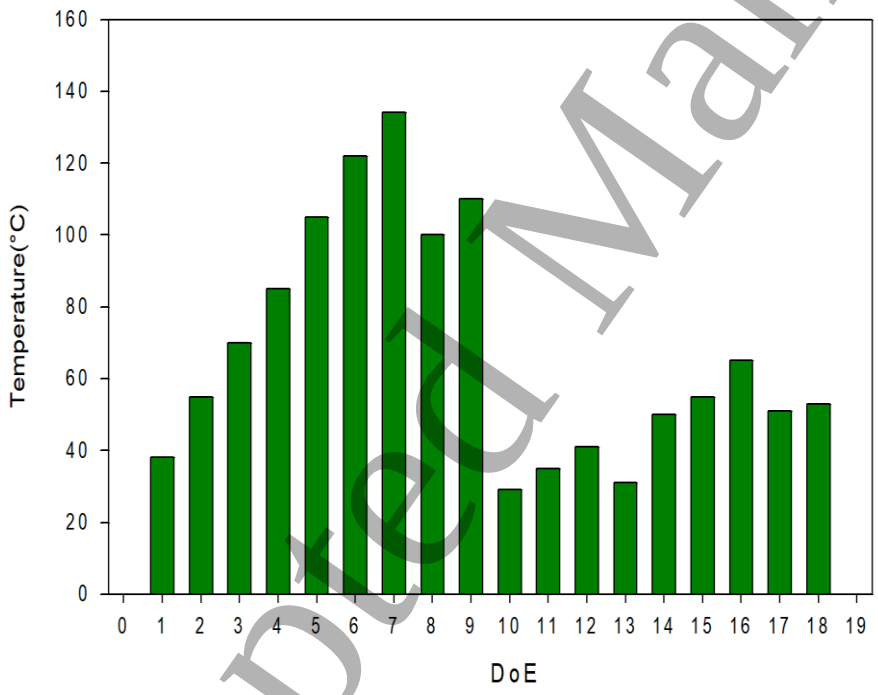


Fig. 8 Experimental T' in accordance with DoE

High temperature leads to more tool wear and not good surface properties. To reduce the temperature LN2 is directly supplied at the interface of tool and workpiece. Due to high heat

Accepted Manuscript

removing rate LN2 is highly efficient in reducing the temperature in short time duration as compared to conventional cooling fluids [7]

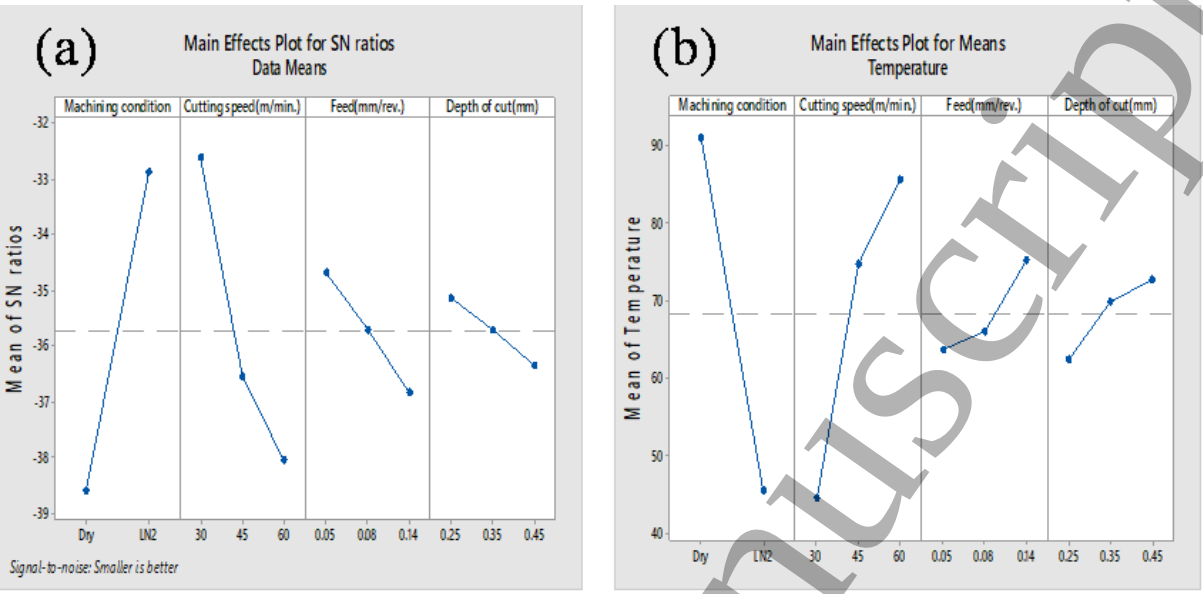


Fig. 9(a) Modification of mean s/n of T'(b) Modification of mean temperature T' at various factors

Fig. 9(a) and Table 6 show modification of mean of s/n of T'. Fig. 9(b) and Table 7 show the modification of mean of T'. On incrementing Vc', Fo' and Dc' temperature at the interface of cutting insert and workpiece increased. Table 6 shows optimum level of temperature at machining (level 2 = LN2), Vc' (level 1 = 30m/min.), Fo' (level 1 = 0.05mm/rev.) and Dc' (level 1 = 0.25mm). The predicted value of temperature has been calculated from Eqs. (2) and (3) is 25.40°C. ANOVA table 8 for response of temperature shows machining condition has the highest effect in percentage of contribution (49.69%), speed has a second higher effect on the percentage of contribution (31.83%), feed has a third higher effect on the percentage of contribution (4.71%) and depth of cut has fourth higher effect in the percentage of contribution (1.48%). P- value is significant for machining condition and speed.

Accepted Manuscript

3.7 Confirmation Experiment: In Taguchi method Confirmation experiments were performed to calculate the difference between actual values and predicted values of response at optimum levels. If the reliability of the condition is assumed to be 95%, then confidence level (CI) is calculated from Eqs. (4) and (5) [56]

$$CI = \sqrt{\left[F(\alpha, 1, fe)Ve\left\{\left(\frac{1}{Ne}\right) + \left(\frac{1}{R}\right)\right\}\right]} \quad (4)$$

$$Ne = \frac{N}{(1+Td)} \quad (5)$$

Where, $F(\alpha,1,f_e)$ is the F-ratio required for 100(1- α) percent confidence level, fe is DOF for error =10, Ve= AdjMs for error, N = total number of experiment (18), R = number of replications for confirmation of experiments (0) and Td = total degree of freedom associated with mean optimum is (7). From standard statistical table, the value of F ratio for α = 0.05 is F(0.05,1,10) = 4.96. Substituting values from ANOVA Table 6 with the respective responses. CI value of surface roughness (Ra') is ±1.53µm, cutting force (Fc') is ±2.09N, flank wear length (Vb') is ±2.42µm and temperature (T') is ± 2.84°C. Experiments have been performed with respective optimum cutting parameters for responses. Surface roughness is 0.80µm, cutting force is 28N, flank wear length is 63µm and temperature is 26°C during machining condition of LN2. The values of responses obtained are within the range of confidence level (CI). The regression models have been generated for studying the relationship between machining parameters and the correlation between independent and dependent parameters. From table 7, this has been observed that R-Sq value is greater than 75% and approaches more than 90% with and a good correlation was obtained between cutting parameters and experimental outputs.

Accepted Manuscript

Table 6 Response table for S/N ratio of surface roughness (Ra'), cutting force (Fc'), flank wear length (Vb') and temperature (T')

Response	Level	M'	Vc'(m/min.)	Fo'(mm/rev.)	Dc'(mm)
Ra'	1	-4.773	-4.350	-2.966	-3.23
	2	**-1.402**	-2.740	**-2.606**	**-2.732**
	3	-	**-2.173**	-3.691	-3.301
	Delta	3.371	2.177	1.085	0.569
	Rank	1	2	3	4
Fc'	1	-38.52	**-32.80**	**-35.08**	-35.52
	2	**-32.70**	-36.33	-35.61	**-35.20**
	3	-	-37.71	-36.14	-36.12
	Delta	5.83	4.91	1.07	0.92
	Rank	1	2	3	4
Vb'	1	-43.58	**-39.60**	**-39.76**	-40.97
	2	**-37.94**	-40.40	-41.26	**-40.65**
	3	-	-42.27	-41.25	-40.65
	Delta	5.64	2.67	1.50	0.32
	Rank	1	2	3	4
T'	1	-38.60	**-32.61**	**-34.68**	**-35.14**
	2	**-32.88**	-36.56	-35.71	-35.72
	3	-	-38.04	-36.83	-36.35
	Delta	5.72	5.43	2.16	1.21
	Rank	1	2	3	4

Accepted Manuscript

Table 7 Response table for means of surface roughness (Ra'), cutting force (Fc'), flank wear length (Vb') and temperature (T')

Response	Level	M'	Vc'(m/min.)	Fo'(mm/rev.)	Dc'(mm)
Ra'	1	1.749	1.685	1.457	1.510
	2	**1.190**	1.403	**1.390**	**1.407**
	3	-	**1.320**	1.562	1.492
	Delta	0.559	0.365	0.172	0.103
	Rank	1	2	3	4
Fc'	1	88.44	**45.90**	**62.73**	63.57
	2	**44.29**	68.82	68.10	**63.25**
	3	-	84.38	68.27	72.28
	Delta	44.16	38.48	5.53	9.03
	Rank	1	2	4	3
Vb'	1	156.56	**97.00**	**103.00**	123.50
	2	**79.56**	113.50	124.83	**112.50**
	3	-	143.67	126.33	118.17
	Delta	77.00	46.67	23.33	11.00
	Rank	1	2	3	4
T'	1	91.00	**44.67**	**63.67**	**62.33**
	2	**45.56**	74.67	66.00	69.83
	3	-	85.50	75.17	72.67
	Delta	45.44	40.83	11.50	10.33
	Rank	1	2	3	4

Accepted Manuscript

Table 8 Analysis of variance for means of surface roughness (Ra'), cutting force (Fc'), flank wear length (Vb') and temperature (T')

Response	Source	DF	Adj SS	Adj MS	F-Value	P-Value	% Cont.
Ra'	M'	1	51.123	51.1229	**47.93**	0.000	**62.42**
	Vc'(m/min.)	2	15.296	7.6481	7.17	0.012	18.67
	Fo'(mm/rev.)	2	3.664	1.8321	1.72	0.228	4.47
	Dc'(mm)	2	1.151	0.5756	0.54	0.599	1.40
	Error	10	10.666	1.0666	-	-	-
	Total	17	81.901	-	-	-	-
Fc'	M'	1	152.781	152.781	**77.27**	0.000	**59.77**
	Vc'(m/min.)	2	77.045	38.522	19.48	0.000	30.14
	Fo'(mm/rev.)	2	3.420	1.710	0.86	0.450	1.34
	Dc'(mm)	2	2.606	1.303	0.66	0.538	1.02
	Error	10	19.772	1.977	-	-	-
	Total	17	255.624	-	-	-	-
Vb'	M'	1	137.470	137.470	**51.40**	0.000	**70.03**
	Vc'(m/min.)	2	23.354	11.677	4.37	0.043	11.89
	Fo'(mm/rev.)	2	8.027	4.013	1.50	0.269	4.09
	Dc'(mm)	2	0.718	0.359	0.13	0.876	0.37

Accepted Manuscript

	Error	10	26.745	2.675	-	-	-
	Total	17	196.314	-	-	-	-
T'	M'	1	147.496	147.496	**40.45**	0.000	**49.69**
	Vc'(m/min.)	2	94.458	47.229	12.95	0.002	31.83
	Fo'(mm/rev.)	2	13.976	6.988	1.92	0.197	4.71
	Dc'(mm)	2	4.397	2.198	0.60	0.566	1.48
	Error	10	36.467	3.647	-	-	-
	Total	17	296.794	-	-	-	-

Table 9 Regression model

Response	Regression Models	R-Sq
Ra'	2.586 - 0.5589 M' - 0.1825Vc'(m/min.) + 0.0525 Fo'(mm/rev.) - 0.0092 Dc' (mm)	83.02%
Fc'	79.9- 44.16 M'+ 19.24 Vc'(m/min.)+ 2.77 Fo'(mm/rev.) + 4.36 Dc'(mm)	84.58%
Vb'	168.9- 77.0 M'+ 23.33Vc'(m/min.)+ 11.67 Fo'(mm/rev.)- 2.67 Dc'(mm)	77.85%
T'	73.8- 45.44 M'+ 20.42 Vc'(m/min.)+ 5.75 Fo'(mm/rev.) + 5.17 Dc'(mm)	81.09%

Accepted Manuscript

3.8 Optimum conditions according to Taguchi

Optimum conditions obtained during experiments for each response i.e. surfaceroughness, cutting force, flankwear length and temperature have been shown in table 11. The machining condition of cryogenic cooling with LN2 has significant effect on values of responses.

Table 10 Optimum condition according to Taguchi

Response	Optimum level of response at control factors
Ra'	M'(Level 2 = LN2), Vc'(level 3=60m/min.), Fo'(level 2= 0.08mm/rev.) and Dc'(level 2= 0.35mm)
Fc'	M'(level 2= LN2), Vc'(level 1= 30m/min.), Fo'(level 1= 0.05mm/rev.) and Dc'(level 2= 0.35mm)
Vb'	M'(level 2=LN2), Vc'(level 1=30m/min.), Fo'(level1=0.05mm/rev.) and Dc'(level 2= 0.35mm)
T'	M'(level 2=LN2), Vc'(level 1=30m/min.), Fo'(level 1= 0.05mm/rev.) and Dc'(level 1= 0.25mm)

3.9 Confirmation test result:

The experiments have been conducted again for validating the predicted values obtained for each response. It has been found that the values obtained after conducting experiment at optimum level are within the predicted range.

Accepted Manuscript

Table 11 Confirmation test result

Response	Predicted value of response	Confidence Level(CI)	Actual value of response at experimental (optimized level)
Ra'	0.75µm	±1.53µm	0.80µm
Fc'	28.01N	±2.09N	28N
Vb'	61.97µm	±2.42µm	63µm
T'	25.40°C	± 2.84°C	26°C

Accepted Manuscript

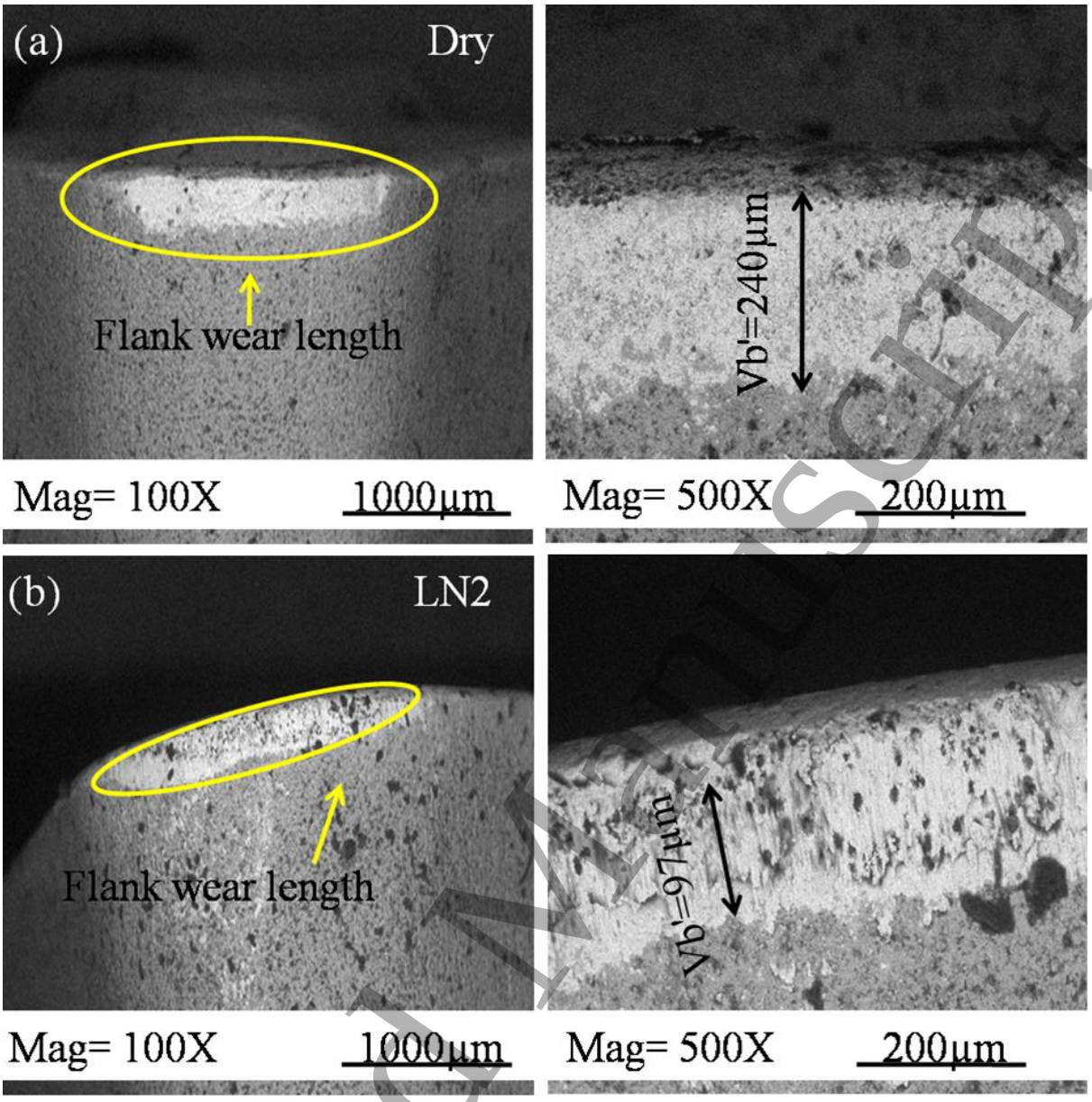


Fig. 10(a) SEM image of the used cutting insert during dry turning at Vc' 60m/min., Fo' 0.14mm/rev. & Dc' 0.25 (b) cryogenic turning with LN2 at Dc '0.35 mm resp.

3.10Surface morphology

The surface generated during dry turning has high surface roughness with some adhesives as shown in SEM image Fig. 11(a). The feed marks are wider in EDS image Fig. 11(b) shows high peaks of iron followed by chromium, carbon and sulphur. The surface generated during cryogenic turning has low surface roughness and narrow feed marks with a clean and smooth

Accepted Manuscript

surface as shown in Fig. 12(a) EDS image Fig. 12(b) show high peaks of iron and other elements. This may be due to low temperature and high pressurised jet takes away adhesive particles from the machined surface.

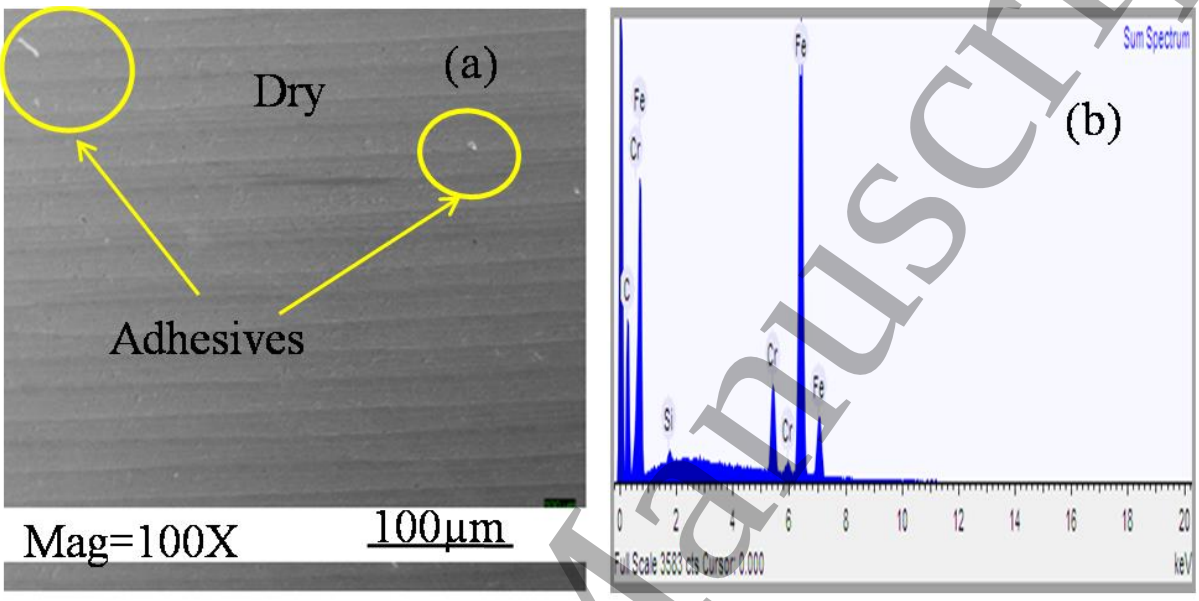


Fig.11 (a) SEM image (b) EDS of the machined surface during dry turning at Vc' 60m/min., Fo' 0.14mm/rev. and Dc' 0.25mm.

Accepted Manuscript

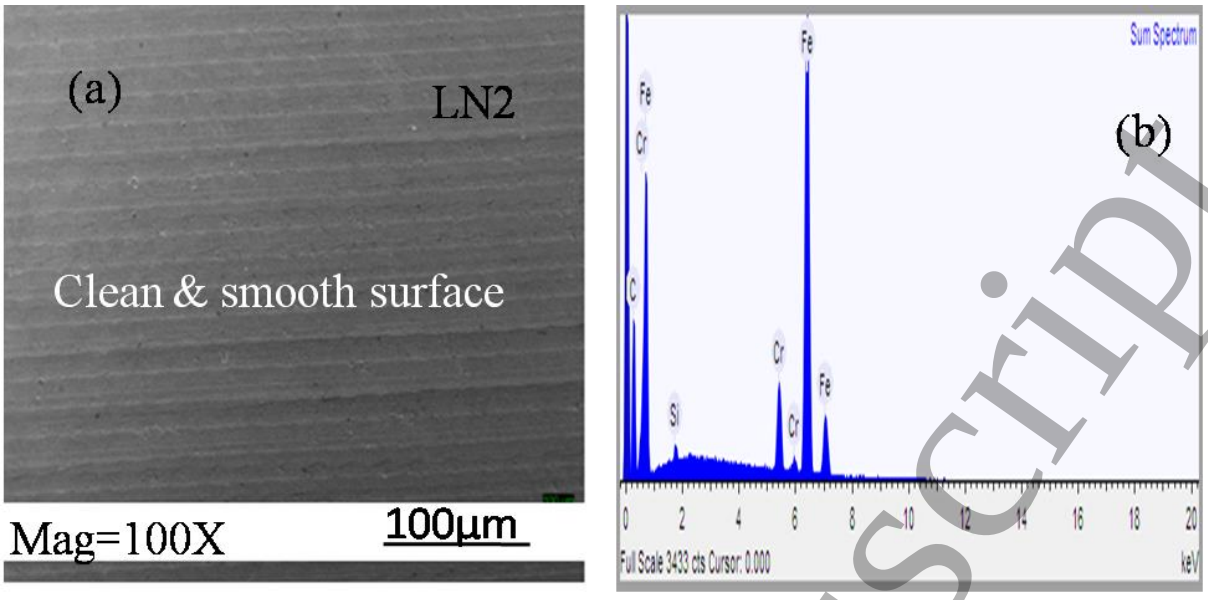


Fig.12 (a) SEM image (b) EDS of the machined surface during cryogenic turning with LN2 at Vc' 60m/min., Fo' 0.14mm/rev. and dc' 0.35 mm

Conclusions:

In the presented research work, the experiments have been performed with titanium nitrate(TiN) coated cutting insert on AISI D3 steel alloy according to Taguchi L_{18} orthogonal array in machining condition of dry and LN2 direct supply at the interface of cutting tool and workpiece. The optimum level of machining parameters and condition have been calculated for responses of machinability characteristics like surface roughness, cutting force, flank wear length and temperature. This supported by ANOVA for calculating the effect of the contribution of control factors. Confirmation tests have been performed and compared with predicted response values. The conclusions based on this study are in the following manner:

1. The optimum level of surface roughness at M' (level 2 = LN2), Vc' (level 3 = 60m/min.), Fo' (level 2 = 0.08mm/rev.) and Dc' (level 2 = 0.35mm).The value of surface roughness is predicted as 0.75µm.

Accepted Manuscript

2. Analysis for the response of surface roughness shows machining condition has the highest effect in the percentage of contribution (62.42%), CI value of surface roughness (Ra') is ±1.53µm. The value of surface roughness after performing the experiment at an optimized control factor is 0.80µm between the predicted range. The R-sq of the model is 83.02%.
3. The optimum level of cutting force at M' (level 2 = LN2), Vc' (level 1 = 30m/min.), Fo' (level 1 = 0.05mm/rev.) and Dc' (level 2 = 0.35mm).The value of cutting force predicted as 28.01N.
4. Analysis for the response of cutting force shows machining condition has the highest effect on the percentage of contribution (59.77%), CI value of cutting force (Fc') is ±2.09N. The value of cutting force after performing the experiment at an optimized control factor is 28N between the predicted range. The R-sq of the model is 84.58%.
5. The optimum level of flank wear length at M' (level 2 = LN2), Vc' (level 1 = 30m/min.), Fo' (level 1 = 0.05mm/rev.) and Dc' (level 2 = 0.35mm).The value of flank wear length is predicted as 61.97µm.
6. Analysis for the response of flank wear length shows machining condition has highest effect on the percentage of contribution (70.03%), CI value of flank wear length (Vb') is ±2.42µm. The value of flank wear length after performing the experiment at an optimized control factor is 63µm. The R-sq of the model is 77.85%.
7. The optimum level of temperature at M' (level 2 = LN2), Vc' (level 1 = 30m/min.), Fo' (level 1 = 0.05mm/rev.) and Dc '(level 1 = 0.25mm). The value of temperature predicted as 25.40°C.
8. Analysis for response of temperature shows machining condition has the highest effect on the percentage of contribution (49.69%), C I value of temperature (T') is ±

Accepted Manuscript

2.84°C. The value of temperature after performing the experiment at optimized control factor is 26°C between the predicted range. The R-sq of the model is 81.09%.

9. Confirmation tests prove that predicted result values are very close to experimental values at optimized control factors of each response.
10. Liquid nitrogen is better feasible coolant available which is eco-friendly and operator's health friendly.

Acknowledgements: Authors are highly thankful to workshop and laboratories facilities shared by Delhi Technological University and Indian Institute of Technology, Delhi (India)

References:

1. Hong S Y, Broomer M (2000) Economical and ecological cryogenic machining of AISI 304 austenitic stainless steel, Clean Technology Environment Policy 2(3): 157-166.
2. Debnath S, Reddy M M, Yi Q S (2014) Environmental friendly cutting fluids and cooling techniques in machining: A review. Journal of Clean and Production 83: 33-47.
3. Byrne G, Scholta E (1993) Environmentally clean machining processes - a strategic approach. CRIP Annals- Manufacturing Technology 42(1): 471-474.
4. Hong S Y, Zaho (1999) Thermal aspects, material considerations and cooling strategies in cryogenic machining. Clean Production Processes1: 107-116.
5. Boubekri N, Foster P R (2015) A technology enabler for green machining: minimum quantity lubrication (MQL). Journal of Manufacturing & management, 212(5): 556-566.

Accepted Manuscript

6. Pereira O, Rodriguez A, Fernandez-A A I, Barreiro J, Lopezde Lacalle L N (2016) Cryogenic and minimum quantity lubrication for eco-friendly turning of AISI 304. Journal of Clean Production 139: 440-449.
7. Shokran A, Dhokia V, Newman S T (2012) Environmentally conscious machining of difficult-to- machine materials with regard to cutting fluids, International Journal of Machine Tools Manufacturing 57: 83-101.
8. Pusavec F, Kramar D, Krajnik P, Kopac J (2010) Transitioning to sustainable production-Part-II: Evaluation of sustainable machining technologies. Journal Clean Production1 18(12): 1211-1221.
9. Pusavec F, Deshpande A, Yang S (2014) Sustainable machining of high temperature Nickel alloy-Inconel 718: Part-I Predictive performance models. Journal of Clean Production 81: 225-269.
10. Tirelli, Chiappini E, Santro M (2014) Economical comparison of cryogenic Vs traditional turning of Ti-6Al-4V:A case study, Key Eng Material (651-653): 1204-1210.
11. Yildiz Y, Nalbant M (2008) A review of cryogenic cooling in machining processes. International Journal of Machine Tools & Manufacture 48: 947-964.
12. Chinchanikar S, Choudhary S K (2015) Machining of hardened steel- Experimental investigations, performance modelling and cooling techniques: A review. International Journal of Machine Tools & Manufacture 89: 95-109.
13. Sharma V S, Dogra M, Suri N M (2009) Cooling techniques for improved productivity in turning. International Journal of Machine Tool & Manufacture 49: 435-453.

Accepted Manuscript

14. Gupta M K, Sood P (2017) Machining comparison of aerospace materials considering minimum quantity cutting fluid: a clean and green approach. Proceedings IMechE, Part C: Journal of Mechanical Engineering Science 231: 1445-1464.
15. Liew P J, Shaaroni A, Nor, Sidik C A, Yan J (2017) An Overview of current status of cutting fluids and cooling techniques of turning hard steel. International Journal of Heat and Mass Transfer 114: 380-394.
16. Yang W H, Tarang Y S (1998) Design optimization of cutting parameters for turning based on Taguchi method. Journal of Materials Processing Technology84: 122-129.
17. Thirumalai R, Senthilkumar J S, Selvarani P, Ramesh S (2012) Machining characterstics of Inconel 718 under several cutting conditions based on Taguchi method. Proceedings of IMeche Part C, Journal of Mechanical Engineering Science 227(9): 1889-1897.
18. Shokrani A, Dhokia V, Newman S T (2016) Investigation of the effects of cryogenic machining on surface integrity in CNC end milling of Ti-6Al-4V titanium alloy. Journal of Manufacturing Processes 2: 172-179.
19. Shokarani A, Dhokia V, Newman S T (2016) Energy conscious cryogenic machining of Ti-6Al-4v titanium alloy. Proc IMechE Part B: Journal Engineering Manufacture: 1-17.
20. Hasclik A, Caydas U (2008) Optimization of turning parameters for surface roughness and tool life based on Taguchi method. International Journal of Advanced Manufacturing Technology 38: 896-903.
21. Mia M, Prithbey, Dey R, Hossain M S, Arafat Md T, Asaduzzaman Md, Ullah Md. S, Zobaer S M T (2018) Taguchi S/N based optimization of machining parameters for surface roughness, too wear and material removal rate in hard turning under MQL cutting condition. Measurement 122: 380-191.

Accepted Manuscript

22. Zerti O, Yallease M A, Khettabi R, Chaoui K, Marbrouki T (2017) Design optimization for minimum technological parameters when dry turning of AISI D3 steel using Taguchi method. International Journal of Advanced Manufacturing Technology 89: 1915-1934.
23. Durakbasa M N, Akdogan A, Vanli A S, Bulutsuz A G (2015) Optimization of end milling parameters and determination of the effects of edge profile for high surface quality of AISI H13 steel by using precise and fast measurements. Measurement 68: 92-99.
24. Gunay M, Yucel E (2013) Application of Taguchi method for determining optimum surface roughness in turning of high-alloy white cast iron. Measurement 46: 913-919.
25. Rath D, Panda S, Pal K (2018) Prediction of surface quality using chip morphology with nodal temperature signatures in hard turning of AISI D3. Materials today 5: 12368-12375.
26. Aggarwal A, Singh H, Kumar P, Singh M (2008) Optimization of multiple quality characteristics for CNC turning under cryogenic cutting environment using desirability function. Journal of materials processing technology 205: 42-50.
27. Mandal N, Doloi B, Mondal B, Das R (2011) Optimization of flank wear using Zirconia toughned aluminia (ZTA) cutting tool: Taguchi method and Regression analysis. Measurement 44: 2129-2155.
28. Nilrundra M, Doloi B, Mondal B (2016) Surface roughness predication model using zirconica toughned aluminia (ZTA): Taguchi method and regression analysis. Journal of Institution of Engineers (India) Series C97 (1): 77-84.
29. Gupta M K, Sood P K (2016) Optimizing multi-characteristics in machining of AISI 4340 steel using Taguchi's approach and utility concept. Journal of Institution of Engineers(India) Series C,97(1): 63-69.

Accepted Manuscript

30. Kumar K V B S Kalyan, Choudhury S K (2008) Investigation of tool wear and cutting force in cryogenic machining using design of experiments. Journal of material processing technology 203: 95-101.

31. Dureja J S, Singh R, Bhatti M (2014) Optimization flank wear and surface roughness during hard turning of AISI D3 steel by Taguchi and RSM methods. Production & Manufacturing Research 2: 767-783.

32. Dr. Vijay K M, Kiran B J, Rudresha N (2018) Optimization of machining parameters in CNC turning of stainless steel(EN19) by Taguchi's orthogonal array experiments. Materials Today Proceedings 5: 11395-11407.

33. Viswanathan R, Ramesh S, Subburam V (2018) Measurement and optimization of performance characteristics in turning of Mg alloy under dry and MQL conditions. Measurement 120: 107-113

34. Yilmaz B, Karabulut S, Gullu A (2018) Performance of analysis of new chip breaker for efficient machining of Inconel 718 and optimization of the cutting parameters. Journal of Manufacturing Processes 32: 553-563.

35. Zebia W, Kowlczyk R (2015) Estimating the effect of cutting data on surfaceroughness and cutting force during WC-CO turning with PCD tool using Taguchi design and ANOVA analysis. International Journal of Advanced Manufacturing Technology 77: 2241-2256.

36. Sun S, Brandt M, Dargush M S (2010) Machining Ti-6Al-4V alloy with cryogenic compressed air cooling. International Journal of Machine Tools & Manufacture 50: 933-942.

37. Manimaran G, Kumar M P, Venkatasamy R (2014) Influence of cryogenic cooling on surface grinding. Cryogenics 59: 76-83.

Accepted Manuscript

38. Manimaran G, Pradeep M (2013) Effect of cryogenic cooling and sol-gel alumina wheel on grinding performance of AISI 316 stainless steel. Archives of civil and mechanical engineering 13: 304-312.

39. Bermingham M J, Kirsch J, Sun S, Palanisamy S, Dargusch M S (2011) New observations on tool life, cutting forces and chip morphology in cryogenic machining Ti6AL4V. International journal of machine tools & manufacture 51: 500-511.

40. Fredj N, Habibsidhon, Braham C (2006) Ground surface improvements of the austentic stainless steel AISI 304 using cryogenic cooling Surface & Coatings technology 200: 4846-4860.

41. Dhananchnezian M, Kumar M P (2011) Cryogenic turning of the Ti-6Al-4V alloy with modified cutting tool inserts. Cryogenics 51: 34-40.

42. Sivaiah P, Charkardhar D (2018) Effect of coolant on turning performance characteristics during machining of 17-4 PH stainless steel: A comparison with MQL, wet, dry machining. CIRP Journal of manufacturing science and technology 21: 86-96.

43. Sartori S, Ghiotti A, Bruschi S (2017) Hybrid lubricating/ cooling strategies to reduce the tool wear in finishing turning of difficult- to- cut alloys. Wear 376-377: 107-114.

44. Dhar N R,Paul S,Chattopadhyay A B (2002) Machining of AISI 4140 steel under cryogenic cooling – toolwear, surfaceroughness and dimensional deviation. Journal of Material processes technology 123: 483-489.

45. Dhar N R, Kishore S V N, Paul S, Chattopadhay A B (2002) Effects of cryogenic cooling on chips and cutting forces in turning AISI 1040 and 4320 steel. Proceedings of IMeche Part-B, Journal of Engineering & Manufacure 216: 713-724.

Accepted Manuscript

46. Dhar N R, Kamruzzaman M (2007) Cutting temperature, tool wear, surface roughness and dimensional deviation in turning AISI-4037 steel under cryogenic condition. International Journal of Machine Tools & Manufacturing 47: 754-759.

47. Dhar N R,Paul S, Chattopadhyay A B (2002) The influence of cryogenic cooling on tool wear, dimensional accuracy and surfacefinish in turning AISI 1040 and E4340C steels.Wear 249: 932-942

48. Hong S Y, Dong Y, Jeong W (2001) Friction and cutting forces in cryogenic machining of Ti-6Al-4V. International Journal of Machine Tools & Manufacturing 41(15): 2271-2285.

49. Hong S Y, Markus I, Jeong W (2001) New cooling approach and tool life improvement in cryogenic machining of titanium alloy. International Journal of Machine Tools & Manufacturing 41(15): 2245-2260.

50. Mia M, Gupta M K, Lozano J A, Carou D,Pimenov D Yu, Krolczyk G, Khan A M, Dhar N R (2018) Multi objective optimization and life cycle assessment of eco-friendly cryogenic assisted turning of Ti-6Al-4V. Journal of Cleaner Production doi: 10.106/j.clepro.2018.10.334.

51. Mia M, Dhar N R (2017) Influence of single and dual cryogenic jets on machinability characteristics in turning of Ti-6Al-4V, Proc IMechE Part B: J Engineering Manufacture 1-16. doi: 10.1177/0954405417737581

52. Strano M, Chiappini E, Tirelli S, Albertelli P, Monno M (2013) Comparison of Ti6Al4V machining forces and tool life for cryogenic versus conventional cooling. Proc IMechE Part B: Journal Engineering Manufacture 227(9): 1403-1408.

53. Diniz1 A E, Machado A R, Correa J G (2016) Tool wear mechanisms in the machining of steels and stainless steels. International Journal Advanced Manufacturing Technology 87: 3157-3168.

Accepted Manuscript

54. Venugopal K A, Tawade R, Prashanth P G, Paul S, Chattopadhyay A B (2003)Turning of titanium alloy with TiB2-coated carbides under cryogenic cooling. Proc. IMechE Part B: J. Engineering Manufacture 217: 1697-1707.

55. Lal D M, Renganarayanan S, Kalanidhi A (2001) Cryogenic treatment to augment wear resistance of tool and die steels. Cryogenics 41: 149-155.

56. Khan A, Maity K (2017) Comparative study of some machinability aspects in turning of pure titanium with untreated and cryogenically treated carbide inserts. Journal of Manufacturing Processes 28: 272-284.

57. Chetan, Ghosh S, Rao P V (2017) Performance evaluation of deep cryogenic processed carbide inserts during dry turning of Nimonic 90 aerospace grade alloy. Tribology International 115:3 97-408.

58. Hakim M A EI, Abad M D, Abdelhameed M M, Shalaby M A, Veldhuis S C (2011) Wear behaviour of some cutting tool materials in hard turning of HSS. Tribology International 44: 1174-1181.

59. Ross P J (1996) Taguchi Techniques for quality engineering, Newyork, McGraw-Hill: 181-196

60. M Vasu , Nayaka H S (2018) Investigation of machinability characteristics on EN47 steel for cutting force and tool wear using optimization technique. Materials Research Express, 5(6), 066501. doi:10.1088/2053-1591/aac67f

61. Bruschi S, Pezzato L, Ghiotti A, Dabala M, Bertolini R (2019) Effectiveness of using low-temperature coolants in machining to enhance durability of AISI 316L stainless steel for reusable biomedical devices. Journal of Manufacturing Processes 39: 295-304.

62. Jadhav P, Kumar S, Bongale A K, Khedkar N (2019) Influence of deep cryogenic cooling on tool wear and surface roughness of coated tungsten carbide inserts using

Accepted Manuscript

statistical techniques, Material Research Express,6(7),076517, doi:10.1088/2053-1591/ab11b8.

63. Fuat Kara (2017) Taguchi optimization of surface roughness and flank wear during the turning of DIN 1.2344 tool steel, Materials Testing 59 (10), 903-908.
64. Fuat Kara, Burak Öztürk (2018) Comparison and optimization of PVD and CVD method on surface roughness and flank wear in hard-machining of DIN 1.2738 mold steel, Sensor Review, 39(1), 24-33, doi/org/10.1108/SR-12-2017-0266.
65. Fuat Kara, Kubilay Aslantas, Adem Cicek (2016), Prediction of cutting temperature in orthogonal machining of AISI316L using artificial neural network" Applied Soft Computing 38, 64–74.
66. **G**upta M K, Singh G, Sood P K (2015) Modelling and optimization of tool wear in machining of EN 24 steel using Taguchi approach, Journal of Institution of Engineers(India) Series C, doi10.007/s40032-015-0175-z

Accepted Manuscript

ISSN: 2393-8528

Contents lists available at www.ijicse.in

International Journal of Innovative Computer Science & Engineering

Volume 6 Issue 2; March-April 2019; Page No. 01-07

EFFECTING PARAMETERS OF BIOGAS PRODUCTION

Ashok Maan[1], Pradeep Kumar Meena[2]

[1]*Department of Mechanical Engineering, Suresh Gyan Vihar University, Jaipur*

[2]*Department of Mechanical Engineering, Delhi Technological University*

ashokmaangvit@gmail.com[1]

ABSTRACT

In this paper we are examining the effective parameters of biogas production. First day we used feed stock rate 150 kg cow dung along with 160 litr water and 5th day again we used 160 kg cow dung along with 250 litr water. After successfully production of methanogenic bacteria we used 10 liter, 13 litr, 20 litr algae at different time of period. We used % of solid content, % of volatile solid, % of ash content, and % of moisture content in feed stock. pH value of different wastage can be calculated by pH meter. Slurry temperature, Volume w.r.t days calculated from 11th days to 40th days.

Keywords- Effecting Parameters, pH value, Feed stock, Ash and moisture content.

1. Introduction:-

The risk of decay of petroleum products because of constant utilize, environmental change in atmosphere conditions and look for interchange assets focuses towards bio-energizes. So utilization of bio-powers with or without petroleum product is a superior method to decrease these kinds of admonitions in the field of vitality and condition. India is in best five nations for utilization of bio-fills. About 2% of aggregate vitality utilization is created from bio-fills in India. Presently days, center is around second era bio-fills because of different disadvantages of original bio-energizes. These second era bio-fills are gotten from lignocelluloses matter or green growth. This undertaking centers around improvement of a green growth based biogas framework. Biogas from green growth is an option in contrast to customary petroleum derivatives and a complementation to first era bio-energizes. In the wake of gathering, vitality transformation process for green growth biomass to biogas will be engaged. Anaerobic absorption process is the most ideal approach to create biogas from green growth biomass on the grounds that different procedures are vitality concentrated or less productive.

The wide goal of this undertaking is to build up a biogas framework for green growth. In this undertaking, center will associate with advancement of an anaerobic digester to expand biogas creation through anaerobic assimilation process by thinking about various influencing parameters. To do contextual investigations of different existing biogas plants with a mean to comprehend the preparing, properties and amounts of crude materials included and parameters which influence the anaerobic assimilation process and nature of biogas delivered. To configuration, create and setting up of anaerobic digester for generation of biogas from green growth. To investigate the execution of anaerobic digester for age of biogas from green growth.

2. Literature review:-

The green growth biomass contains starches, fats and proteins. The green growth species which have just single cell is called microalgae. Green growth are delivered in open lakes and in shut photograph bioreactors [1]. The span of microalgae is

*Corresponding author: Ashok Maan

little and can be seen with magnifying lens just [2].Main types of microalgae are chlorella and spirulina, which are utilized for the generation of bio-fuel [3].The development time of green growth is short, and mass of green growth will be twofold in one to four days. Green growth which have least 30%lipids are increasingly appropriate for bio-fuel creation [4]. Starch in the green growth is the primary constituent for generation of biogas, which is created by the change of sugar. Sugar is created by photosynthesis process. [5]. CO2 is required for development of green growth and it is given in the water. Sharpness and alkalinity of the water, with some physical conditions influence the development of green growth [6]. Open lakes of raceway type can create green growth biomass in the scope of 10 to 25 gram for every square meter [7]. Biogas can be created from wet biomass of green growth by anaerobic absorption process. At present numerous investigations have been completed on the anaerobic processing of green growth [8]. Anaerobic assimilation of green growth is a test because of its thick cell dividers. Thickness of cell dividers of green growth is 0.1 to 0.3 micrometer [9].

3. Methodology and Result-

3.1 Calculation for anaerobic digester

Under mesophilic conditions and Retention time=30 days

Considering Volume of digester = 1000 liter with a drifting drum limit of 500 liter. For this a water tank of 1000 liter and 500 liter limit can be utilized.

Slurry volume= volume of digester/Retention to time = 1000/6 = 170 liter =1000-170=830 liter

(1/6 part of digester is utilized as dead territory and rest for slurry)

Slurry volume= 830 liter

Gas storage= 170 liter

Rate of feedstock=Active digester volume/Retention time = .830/30=.02785 m3/day

From Bus well condition, Algae: water=1:1.3

Mass of algae= (1/2.3)*.02785=.0121m3/day

Thickness of green growth =800 kg per shape meter =.0121*800=9.7Kg/day

Thickness of water=1000Kg/day

Mass of Water=(1.3/2.3)*.02785=.01575m3 /day =.01575*1000=15.75 kg/day

Volume of slurry=9.7+15.75=25.45 Kg/day

Add up to slurry after 30 days=25.45*30=763 kg

Organization of lake green growth is appeared in table 1.

Table 1: Pond algae analysis:-

Moisture	% air dried volatile matter	Fixed carbon	Ash	% dry ash free (VM)
4.01	64.60	6.4	25	90.99

Unstable solid= 64.9%

Stacking rate:- Total VS produced=amount of green growth feedstock*VS% = 9.7*.65=6.3 kg/day

Expect 60% VS expulsion at 30 days SRT under mesophilic conditions

Versus removed= VS loading*VS evacuation effectiveness = 6.3*.60 Kg/day= 3.8 kg/day

One kg VS produce=0.271m3 biogas under perfect conditions 3.8 Kg VS deliver =1.025m^3 biogas

Thickness of biogas=1.15 Kg/m^3

Mass of biogas produced=1.15*1.025 =1.18 Kg every day

Add up to strong waste from digester

(Add up to solids-Volatile solids)+ (Total unstable VS evacuation) (9.7-6.3) + (6.3-3.8)= 5.9 kg every day

3.2 Steering framework:-

Thickness of green growth is 800 kg for every cubic meter. So a legitimate blending of green growth biomass with slurry in the digester is

fundamental. For this instigator is utilized. This can be physically, mechanically or electrically. In this digester, outspread sort instigator which is worked physically is utilized. Spiral sort instigator is utilized for legitimate blending. By legitimate blending microorganism in the digester have more region for processing. So processing rate of feedstock is expanded. Fomenter has four sharp edges at an edge of 45°. At this position edge of fomenter connected greatest power to the slurry. This instigator is settled in the drifting drum. By turning the drifting drum blend the crisp feed stock is blended with slurry.

3.3 Digester Heating:-

For effective creation of biogas, great conditions ought to be kept up in the digester on the grounds that anaerobic absorption is an unpredictable procedure. Execution parameters like strong stacking rate, blending, pH are autonomous of condition. Be that as it may, temperature of digester is influenced by encompassing temperature. So it is most influenced parameter in the assimilation procedure. For a mesophilic digester temperature ought to be kept up at 35-40 degree C.

At that point two conditions ought to be satisfied in the digester:-

- Raising the digester temperature up to mesophilic temperature through a warmth source;
- Maintaining the temperature in digester by ideal protection of digester.
- Two kinds of misfortunes from the digester,
- Heat lost from the assemblage of digester because of encompassing conditions,
- Adding the feedstock in the digester.

Warmth lost from the digester can be lessened by giving the protection.

To bring the temperature up in digester a curl sort of warmth exchanger is utilized. The working liquid in warmth exchanger is water.

Figure 1:

In the wake of setting up of digester, diverse operational and execution parameters, for example, feed rate, nature of feed material, maintenance time, temperature, pH, volume of gas created, and nature of gas are watched. On 18 March, 2018 the digester was begun with 300kg slurry (150kg bovine dung+ 150 liter water). Following five days, 23 March, 2018 slurry in digester was additionally included (160 kg dairy animals manure and 250 liter water) and left for multi week. The nature of biogas delivered was checked with a 10 days hole. The temperature and ascent of skimming drum were noted to gauge volume of biogas delivered on hourly premise.

3.4 Feedstock rate:-

The feed rate of feedstock (cow dung and algae) in the digester is given in table 2.

Table 2: Feedstock rate for digester

1st day	150kg cow dung+160 liter water
5th day	160 kg cow dung + 250 liter water
20th day	10 liter algae
25th day	13 liter algae
30th day	20 liter algae

3.5 Composition of feedstock:-

3.5.1 Solid substance in feedstock: - To decide the strong substance in feedstock, take the example of feedstock. Taken the underlying weight of the example and set in stove for one hour at 105 °C for expelling dampness. Following one hour taken the

heaviness of dried example and ascertain the strong substance.

$$\% \text{ Solid content} = \frac{\text{final weight}}{\text{Initial Weight}} \times 100 \text{---------1}$$

Solid content in cow dung and algae is given in table 3.

Table 3: Solid Content in feedstock

Feedstock	Initial weight(gm)	Final weight(gm)	% Solid content
Fresh cow dung	1.0199	0.4976	48.79
Algae	1.0494	0.2578	24.56

3.5.2 Volatile solid in feedstock:-To determine volatile material one gram sample was put in a muffle furnace at 650°C for 7 minutes. The sample was partially covered in the crucible.

$$\% \text{ Volatile content} = \frac{\text{Initial weight} - \text{final weight}}{\text{Initial Weight}} \times 100$$

---------- (ii)

Volatile matter in cow dung, algae and in digestive material is given in table 4.

Table 4: Volatile Solid in Feedstock

Feedstock	Initial weight(gm)	Final weight(gm)	% Volatile Solid
Fresh cow dung	1.0557	0.1843	82.54
Algae	1.0017	0.2650	73.54
Digestive material	1.0030	0.2021	79.85

3.5.3 Ash content: -To determine ash content placed the sample in muffle furnace at 650°C for 30 minutes.

$$\% \text{ Ash content} = \frac{\text{Final weight}}{\text{Initial Weight}} \times 100 \text{.....(iii)}$$

Ash content in cow dung, algae and in digestive material is given in table 5.

Table 5: Ash Content in Feedstock

Feedstock	Initial weight(gm)	Final weight(gm)	% Ash content
Fresh cow dung	1.1669	0.1485	12.72
Algae	1.1468	0.0664	5.79
Digestive material	1.0464	0.0940	8.98

3.5.4 Moisture content:-To decide dampness in feedstock, squashed the example in powder frame and set in boiler at 105°C in a silica pot for two hours. At that point following two hours, cooled the example and weighted. At that point procedure was rehashed for 3 hours.

$$\% \text{ Moisture content} = \frac{\text{Initial weight of Solid} - \text{Weight of dry solid}}{\text{Initial Weight}} \times 100$$

---------- (12)

Moisture content in cow dung, algae and in digestive material is given in table 6 .

Table 6: Moisture Content in Feedstock

Feedstock	Initial weight(gm)	Final weight(gm)	% Moisture content
Fresh cow dung	1.0199	0.9942	2.52
Algae	1.0640	1.0143	4.67
Digestive material	1.0494	0.9905	5.61

3.5.5 Fixed carbon: 100-(%VM-%Ash-Moisture)

Summary of composition of feed stocks is given in table 7.

Table 7: Composition of feedstock

Feedstock	VM (%)	Ash content (%)	Moisture (%)	Fixed carbon (%)
Cow dung(fresh)	84.34	11.72	2.52	1.42
Algae	73.54	10.82	4.67	10.97
Cow dung(digestive)	79.85	8.98	5.61	5.56

3.6 pH Value:

pH esteem is a proportion of nature of slurry. For small scale living being development in digester pH esteem ought to be over 6 and underneath 8. To quantify the pH of slurry an advanced pH meter was utilized. An example of slurry was taken in a spotless test cylinder to quantify pH. The test of pH meter was plunged in slurry test. It gave the incentive at computerized marker. The estimations of pH are inside the range as given in table 8.

Table 8: pH value of slurry

Parameter	Fresh cow dung slurry	Fresh Algae slurry	Digester slurry
pH	6.56	6.62	6.54

3.7 Slurry temperature, Volume vs. days:-

To gauge the slurry temperature a thermocouple with a computerized pointer was utilized. Perusing was taken from10:00 AM to 5:00 PM. Volume of delivered biogas was estimated with a chart paper which was settled at the gliding drum. Ascending of drum show the generation of biogas. By seeing the stature ascent of the skimming drum, the volume of created gas was determined.

Figure 1: Thermocouple for temperature measurement

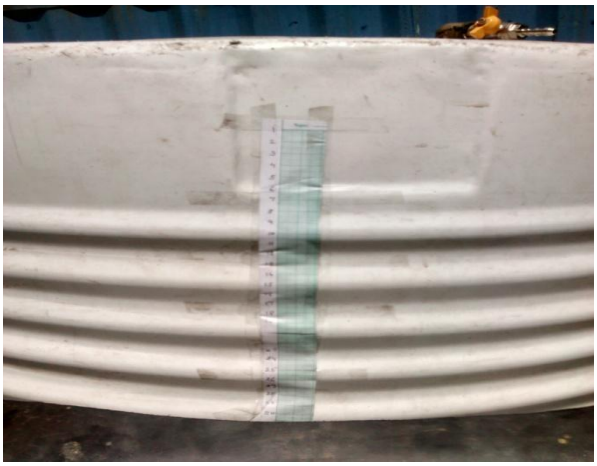


Figure 2: Graph scale to measure rising height of floating drum

Slurry temperature and volume of produced biogas on 11th day is shown in figure 3.

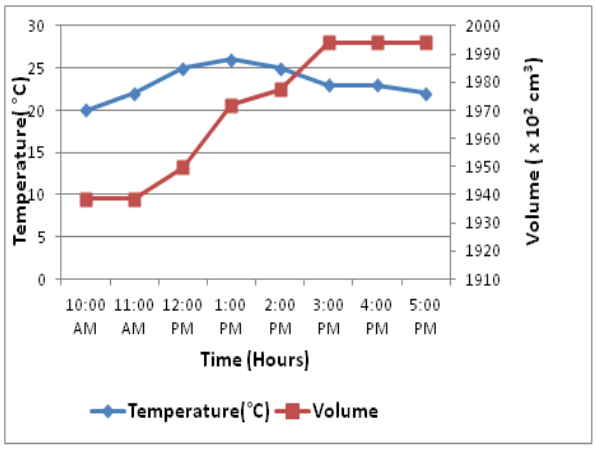


Figure 3: Temperature and volume on 11th day

This diagram demonstrates that slurry temperature is low. This is beginning of mesosphilic temperature go. Volume of biogas created is less. Slurry temperature and volume of created biogas on twelth day is appeared in figure 4.

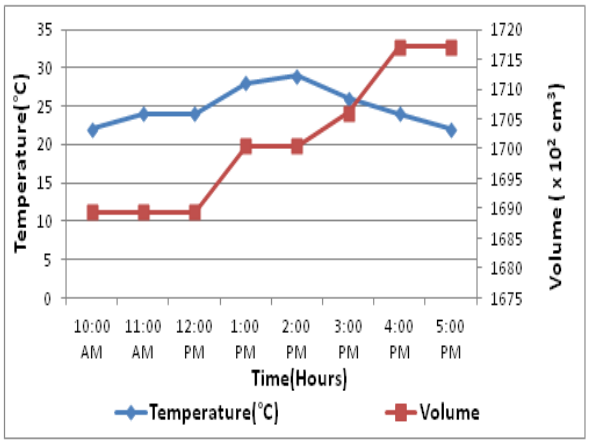


Figure 4: Temperature and volume on 12th day

This diagram demonstrates that as the temperature builds, volume of created biogas likewise increments. Amid morning time temperature is low, volume of biogas is additionally less. As the temperature is at top between 1:00 to 2:00 pm, more volume of biogas is delivered. Slurry temperature and volume of delivered biogas on thirteenth day is appeared in figure 5.

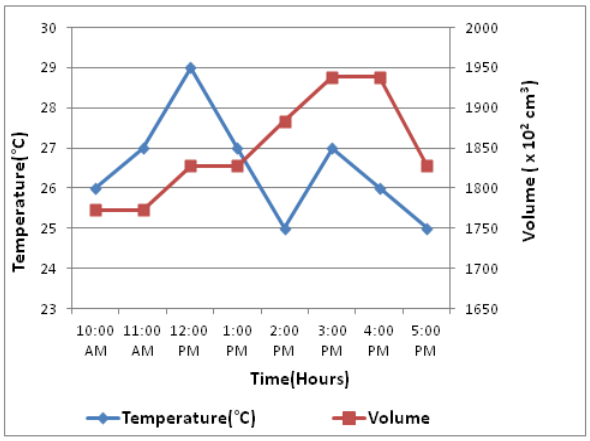


Figure 5: Temperature and volume on 13th day

Similarly we found out volume of production w.r.t temperature from 14th days to 40th days. This graph shows that as the temperature changes, volume of production of biogas also changes. This shows the direct effect on production of biogas. Slurry temperature and volume of produced biogas on 40th day is shown in figure 6.

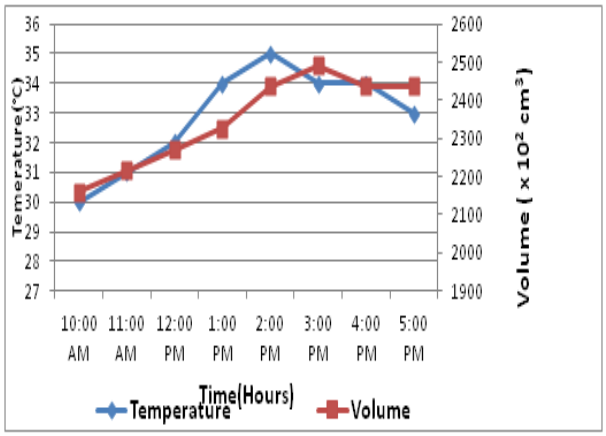


Figure: 6: Temperature and volume on40th day

Digester slurry attains 30 days retention time, feeding in proper ratio and slurry temperature was good. Digester was working in efficient way and production of biogas was increased day by day.

4. Conclusions:-

- Case thinks about demonstrate that biogas can be created from a productive biogas plant by utilizing diverse substrates.
- Results demonstrate that nourishing ought to be appropriate. High stacking rate can influence the biogas generation.

- Retention time assumes an imperative job underway of biogas from a digester. There ought to be least 20 days maintenance time for generation of biogas. Methane focus in biogas created for the most part relies on maintenance time.
- Temperature is a central point for development of smaller scale life forms. Temperature of slurry in the digester ought to in the mesophilic go. Amid summer, temperature is adequate for biogas digester and amid winter, a warmth source is required to keep up the temperature in the digester. Generation of biogas was less during the evening and all the more amid day time. This shows impact of temperature on creation of biogas.
- Methane focus was about 15% following 10 days and about 54% following 30 days. Methane focus was expanded step by step.
- Volume of biogas created was expanded as more maintenance time achieved by digester.
- After bolstering green growth in digester comparable outcomes were seen likewise with dairy animals fertilizer. Green growths are great substrate for biogas digester.

5. Future Scope of Work

1. In the biogas digester

- In this digester about 25% cow excrement is supplanted with green growth. In continuation of this venture 100% dairy animals waste can be supplanted with green growth
- Cow compost and green growth have low C: N proportion. A carbon rich co-substrate can be utilized with green growth.

2. Usage of biogas

- For the usage of biogas in vehicles, scouring is required. A proficient cleaning procedure can be utilized.
- Easy stockpiling and transportation of biogas are required for better usage of biogas.

References

1. Bio fuels from microalgae by Parliament house of science and technology, POSTNOTE NUMBER, 2011
2. Becker. W,Micro, Algae in human and animal nutrition, handbook of micro algae culture. Blackwell oxford, 2004
3. Chisti Y, Biodiesel from microalgae. Biotechnology Advances, 2007
4. Yue Wang,Micro algae as the third generation bio-fuels: production, usage, challenges and prospects ,2013
5. Teodorita Al Seadi, Dominik Rutz, Heinz Prassl, Michael Köttner, Tobias Finsterwalder,Silke Volk, Rainer Janssen, Biogas Handbook published by Southern Denmark, 2008
6. Krogen M and Muller Langer, Review on possible algae biofuel production processes, 2012
7. Poster C and schaub G Micro algae and terrestrial biomass and sources for fuels; Journal of biotechnology, 2009.
8. Goloeke CG, Oswald WJ Gotas HB Anaerobic Digestion of algae, Applied Microbial, 1957
9. Demirbas A and Faith M, Importance of algae oil as a source of biodiesel. Energy Conversion and Management ,2011

Circuits, Systems, and Signal Processing
https://doi.org/10.1007/s00034-019-01195-x

Efficient MDCT Recursive Structure for VLSI Implementation

Pragati Dahiya[1] · **Priyanka Jain**[1]

Received: 16 July 2018 / Revised: 29 June 2019 / Accepted: 1 July 2019
© Springer Science+Business Media, LLC, part of Springer Nature 2019

Abstract
Modified discrete cosine transform (MDCT) is a computationally intensive fundamental processing block for mapping time into frequency in various versions of MPEG audio coding. An efficient algorithm for MDCT computation is a basic requirement to realize low-cost audio codec. This paper introduces a new recursive algorithm and architecture for MDCT. The proposed architecture requires minimum hardware and execution time. Folding of the input sequence is performed multiple times in the subsequent groups which results in a significant reduction in real multiplication and real addition operations. Also, fewer number of execution cycles are required to generate the output through this algorithm. The proposed structure is suitable for parallel VLSI implementation as it is simple, regular and modular.

Keywords Modified discrete cosine transform (MDCT) · Infinite impulse response (IIR) filter structure · Recursive algorithm · Very large scale integrated circuits (VLSI) · MPEG codec

1 Introduction

Time domain aliasing cancelation (TDAC) filter bank is used for analysis and synthesis of signals in audio codec. It comprises of two operations: First is modified discrete cosine transform (MDCT), and second is windowing process. MDCT is majorly used for compression of digital audio signals in MPEG-1-layer III ISO/IEC11172-3, MP3, MPEG-2 AAC (advanced audio coding), MPEG-4 audio ISO/IEC 14496-2, AAC (HE-AAC), Sony ATRAC (adaptive transform acoustics coding), AT&T perceptual audio coder (PAC), Dolby Digital, patent and royalty-free Ogg Vorbis codec [1]. Other applications of MDCT are voice encoding [18], audio watermarking [28], noise reduction

✉ Priyanka Jain
priyajain2000@rediffmail.com

Pragati Dahiya
pragatidahiya@gmail.com

[1] Department of Electronic and Communication Engineering, Delhi Technological University, Bawana Road, Delhi, India

Published online: 18 July 2019

[24], etc. MDCT is a lapped transform in which adjacent data block overlaps by $N/2$ samples. The combination of MDCT and MDST forms basis vector of modulated complex lapped transform (MCLT) [1]. Unlike other discrete trigonometric transforms, forward MDCT generates $N/2$ frequency coefficients for an input sequence of length N. Many algorithms for efficient computation of MDCT have been reported in the literature [3, 19–21, 25–27]. These algorithms comprise of fast and recursive algorithms. One of the major methods for the development of the fast algorithms is by performing sparse matrix factorization of the respective transform matrix. Other methods include polynomial algebras, displacement approach and polynomial division in matrix form [22]. MDCT equation is reduced to other transform formulations like DFT (discrete Fourier transform), DST-II (discrete sine transform)/DCT-II (discrete cosine transform), DCT/DST-IV of lower size [1, 2, 11, 25–27, 30]. Fast algorithms compute all transform coefficients in parallel and are suitable for high processing speed applications, whereas for applications where some critical coefficients are to be calculated and simple structure is a prerequisite condition, recursive structures are used. Implementation of fast algorithm for varying block sizes requires different structures and computation of MDCT using DFT or DCT involves structural irregularity which results in more number of multipliers and adders, thereby increasing the hardware implementation cost. In recursive algorithms, the data are given as input serially and all the coefficients are calculated recursively within the same recursive structure. In the case of recursive algorithms, a simple recursive filter structure computes all the coefficients of variable or general length MDCT. The recursive structures are simple with low control complexity and can be easily implemented in VLSI. Chiang and Liu [6] proposed a recursive algorithm for MDCT/IMDCT which is suitable for online computation scheme. Nikolajevic and Fettweis [19–21] have suggested MDCT/IMDCT algorithm by applying different recurrence formulas and realized their corresponding structures. Clenshaw's recurrence formula is used to derive recursive formulas for MDCT and IMDCT in [20]. However, the significant drawback in these works is that N execution cycles are consumed to compute a transform coefficient for N-length MDCT. To reduce this latency and computational complexity of these recursive algorithms, Chen and others [5] proposed MDCT and IMDCT algorithms which compute transform coefficients in about 72% computation cycles compared to the general length structures. Even though the data throughput (DTPT) is twice in Chen [5] structure, hardware requirement is more compared to the variable length structure. Lai et al. [12–16] presented unified recursive architecture to compute MDCT/IMDCT with low-cost and high-accuracy design. In this paper, recursive algorithm of MDCT is proposed and its corresponding structure is realized through infinite impulse response (IIR) filter structure, which requires fewer execution cycles. Also, a lesser number of hardware components are required for its implementation in comparison of previously reported structures in the literature. After folding the MDCT kernel as given in [5] and based on the concept used in algorithm proposed in [4], MDCT kernel is transformed into sinusoidal recursive formula and its recursive structure is realized for an input block of size $N = 2^r$, where $r > 1$. The total number of real multiplications and additions required in proposed MDCT algorithm is lesser compared to the previously reported algorithms [5, 6, 12, 13, 19–21]. Rest of the paper is organized as follows: Sect. 2 describes the derivation of proposed MDCT algorithm and their realization through IIR filter structure; in

Sect. 3, hardware complexity and computational complexity of the suggested structure are compared with other recursive structures available in the literature, and finally, the conclusions of the proposed work are summarized in Sect. 4.

2 Derivation of Recursive Algorithm for MDCT

The MDCT of input sequence $x(i)$, $i = 0, 1, 2, \ldots, N-1$, where $N = 2^r$ and $r > 1$ is given by the following relation:

$$Y(k) = \sum_{i=0}^{N-1} x(i) \cos\left[\frac{(2k+1)\left(2i+1+\frac{N}{2}\right)\pi}{2N}\right], \quad k = 0, 1, 2, \ldots, \frac{N}{2}-1 \tag{1}$$

Summation in Eq. (1) is decomposed into two summation terms

$$\begin{aligned} Y(k) = &\sum_{i=0}^{\frac{3N}{4}-1} x(i) \cos\left[\frac{(2k+1)\left(2i+1+\frac{N}{2}\right)\pi}{2N}\right] \\ &+ \sum_{i=\frac{3N}{4}}^{N-1} x(i) \cos\left[\frac{(2k+1)\left(2i+1+\frac{N}{2}\right)\pi}{2N}\right], \quad k = 0, 1, 2, \ldots, \frac{N}{2}-1 \end{aligned} \tag{2}$$

Substituting $i' = i + \frac{N}{4}$ in first summation and $i'' = i - \frac{3N}{4}$ in the second summation and replacing the dummy variables i' and i'' by variable i, we get

$$\begin{aligned} Y(k) = &\sum_{i=\frac{N}{4}}^{N-1} x\left(i-\frac{N}{4}\right) \cos\left[\frac{(2k+1)(2i+1)\pi}{2N}\right] \\ &- \sum_{i=0}^{\frac{N}{4}-1} x\left(i+\frac{3N}{4}\right) \cos\left[\frac{(2k+1)(2i+1)\pi}{2N}\right], \quad k = 0, 1, 2, \ldots, \frac{N}{2}-1 \end{aligned} \tag{3}$$

$$\text{Let } \; g(i) = x\left(i-\frac{N}{4}\right), \quad i = \frac{N}{4}, \ldots N-1 \tag{4}$$

$$g(i) = -x\left(i+\frac{3N}{4}\right), \quad i = 0, \ldots, \frac{N}{4}-1 \tag{5}$$

Equation (3) can be written as

$$Y(k) = \sum_{i=0}^{N-1} g(i) \cos\left[\frac{(2k+1)(2i+1)\pi}{2N}\right], \quad k = 0, 1, 2, \ldots, \frac{N}{2}-1 \tag{6}$$

From Eq. (6), it can be observed that MDCT equation reduces to DCT-IV by rearranging the input sequence order as shown in Eqs. (3)–(5). Recently, Perera et al. [23] have introduced fast and recursive DCT-II and DCT-IV algorithms with low arithmetic count. Many systolic algorithms are also available in the literature [7–10] for DCT-IV computation. These architectures are good VLSI solution as these are based more on communicational complexity than computational one.

Now, the sequence $g(i)$ is folded

$$Y(k) = \sum_{i=0}^{\frac{N}{2}-1} \left[g(i) - g(N-1-i)\right] \cos\left[\frac{(2k+1)(2i+1)\pi}{2N}\right], \quad k = 0, 1, 2, \ldots, \frac{N}{2} - 1 \tag{7}$$

$$\cos\left[\frac{(2k+1)(2i+1)\pi}{2N}\right] = \frac{1}{2\cos\left[\frac{(2i+1)\pi}{2N}\right]}\left[\cos\left[\frac{(k+1)(2i+1)\pi}{N}\right] + \cos\left[\frac{(2i+1)k\pi}{N}\right]\right] \tag{8}$$

Using the relation given in Eq. (8) to modify Eq. (7), we obtain

$$\begin{aligned} Y(k) = &\sum_{i=0}^{\frac{N}{2}-1} q_0(i) \cos\left[\frac{(k+1)(2i+1)\pi}{N}\right] \\ &+ \sum_{i=0}^{\frac{N}{2}-1} q_0(i) \cos\left[\frac{(2i+1)k\pi}{N}\right], \quad k = 0, 1, 2, \ldots, \frac{N}{2} - 1 \end{aligned} \tag{9a}$$

where

$$q_0(i) = \frac{1}{2\cos\left[\frac{(2i+1)\pi}{2N}\right]}\{g(i) - g(N-i-1)\} \tag{9b}$$

$$Y(k) = Q(k+1) + Q(k), \quad k = 0, 1, 2, \ldots, \frac{N}{2} - 1 \tag{10}$$

Assume

$$Q(k) = \sum_{i=0}^{\frac{N}{2}-1} q_0(i) \cos\left[\frac{(2i+1)k\pi}{N}\right], \quad k = 0, 1, 2, \ldots, \frac{N}{2} - 1 \tag{11}$$

The term $Q(k)$ is equivalent to even-indexed DCT-II coefficients. The symmetricity of cosine kernels is utilized to compute $Q(k)$ with reduced number of multiplications and additions. $Q(k)$ is categorized into $(r-1)$ number of groups with group number p and a D.C term.

$$\begin{aligned} Q(k) = Q(2^p(2n+1)) = Q(2n+1), Q(2(2n+1)), \ldots, Q\left(2^{(r-2)}(2n+1)\right) \\ \text{and} \quad Q(0), \quad n = 0, 1, 2, \ldots, 2^{(r-p-2)} - 1 \end{aligned} \tag{12}$$

$$\text{Let } p = 0, Q(2n+1) = Q(1), Q(3), Q(5), \ldots, Q\left(2^{(r-1)} - 1\right), \quad n = 0, 1, 2, \ldots, 2^{(r-2)} - 1 \tag{13}$$

$$\text{and } p = 1, Q(2(2n+1)) = Q(2), Q(6), Q(10), \ldots, Q\left(2^{(r-1)} - 2\right), \quad n = 0, 1, 2, \ldots, 2^{(r-3)} - 1 \tag{14}$$

Similarly, all other groups' coefficients are categorized

$$p = r - 2, \quad Q\left(2^{(r-2)}(2n+1)\right) = Q\left(2^{(r-2)}\right), \quad n = 0 \tag{15}$$

No multiplication operations are required for computation of $Q(0)$. It is defined as

$$Q(0) = \left\{ q_{r-2}(0) + q_{r-2}\left(\frac{N}{2^{r-1}} - 1\right) \right\}, \tag{16}$$

Replacing $k = 2^p(2n+1)$ and by folding $q_0(i)$ once, we get

$$\begin{aligned} Q(2^p(2n+1)) = \sum_{i=0}^{\frac{N}{4}-1} & \left\{ q_0(i) + q_0\left(\frac{N}{2} - i - 1\right) \cos\left[2^p(2n+1)\pi\right] \right\} \\ & \cos\left[\frac{(2i+1)2^p(2n+1)\pi}{N}\right], \quad p = 0, 1, 2, \ldots, (r-2) \end{aligned} \tag{17}$$

For $p = 0$, Eq. (17) can be rewritten as:

$$Q((2n+1)) = \sum_{i=0}^{\frac{N}{4}-1} \left\{ q_0(i) - q_0\left(\frac{N}{2} - i - 1\right) \right\} \cos\left[\frac{(2i+1)(2n+1)\pi}{N}\right], \quad p = 0 \tag{18}$$

For $p > 0$, Eq. (17) can be rewritten as:

$$Q(2^p(2n+1)) = \sum_{i=0}^{\frac{N}{4}-1} \left\{ q_p(i) + q_p\left(\frac{N}{2} - i - 1\right) \right\} \cos\left[\frac{(2i+1)2^p(2n+1)\pi}{N}\right], \quad p = 1, 2, 3 \ldots (r-2) \tag{19}$$

In Eqs. (18) and (19), $q_0(i)$ sequence is folded with different sign for $p = 0$ and $p > 0$ group.

Let

$$q_1(i) = q_0(i) + q_0\left(\frac{N}{2} - i - 1\right) \tag{20}$$

Similarly, for $p = 1$ group, $q_1(i)$ is folded

$$\begin{aligned} Q(2^p(2n+1)) = \sum_{i=0}^{\frac{N}{8}-1} & \left\{ q_1(i) + q_1\left(\frac{N}{4} - i - 1\right) \cos\left[2^{p-1}(2n+1)\pi\right] \right\} \\ & \cos\left[\frac{(2i+1)2^p(2n+1)\pi}{N}\right], \quad p = 1, 2, 3 \ldots, (r-2) \end{aligned} \tag{21}$$

Generalized expression of $Q(k)$ is given by:

$$Q(2^p(2n+1)) = \sum_{i=0}^{\frac{N}{2^{p+2}}-1} \left\{ q_p(i) - q_p\left(\frac{N}{2^{p+1}} - i - 1\right) \right\} \cos\left[\frac{(2i+1)2^p(2n+1)\pi}{N}\right], \quad p = 0, 1, 2, \ldots, (r-2) \tag{22a}$$

where

$$q_{p+1}(i) = q_p(i) + q_p\left(\frac{N}{2^{p+1}} - i - 1\right) \tag{22b}$$

Equation (22a) can be written as

$$Q(2^p(2n+1)) = \sum_{i=0}^{\frac{N}{2^{p+2}}-1} w_p(i) \cos\left[\frac{(2i+1)2^p(2n+1)\pi}{N}\right], \quad p = 0, 1, 2, \ldots, (r-2) \tag{23a}$$

where

$$w_p(i) = \left\{ q_p(i) - q_p\left(\frac{N}{2^{p+1}} - i - 1\right) \right\} \tag{23b}$$

Replacing i by $\left(\frac{N}{2^{p+2}} - 1 - i\right)$ in Eq. (23a), we get

$$Q(2^p(2n+1)) = (-1)^n \sum_{i=0}^{\frac{N}{2^{p+2}}-1} w_p\left(\frac{N}{2^{p+2}} - 1 - i\right) \cos\left[\frac{(2n+1)\pi}{2} - (2i+1)\frac{2^p(2n+1)\pi}{N}\right], \quad p = 0, 1, 2, \ldots, (r-2) \tag{24}$$

$$Q(2^p(2n+1)) = (-1)^n \sum_{i=0}^{\frac{N}{2^{p+2}}-1} w_p\left(\frac{N}{2^{p+2}} - 1 - i\right) \sin\left[\left(i + \frac{1}{2}\right)\alpha_{p,n}\right], \quad p = 0, 1, 2, \ldots, (r-2) \tag{25a}$$

$$\text{where} \quad \alpha_{p,n} = \frac{2^{p+1}(2n+1)\pi}{N} \tag{25b}$$

$$Q(2^p(2n+1)) = (-1)^n \sum_{i=0}^{j} w_p(j-i) \sin\left[\left(i + \frac{1}{2}\right)\alpha_{p,n}\right], \quad p = 0, 1, 2, \ldots, (r-2) \tag{26}$$

Assume

$$H_{jn}^P(k) = (-1)^n \sum_{i=0}^{j} w_p(j-i) \sin\left[\left(i + \frac{1}{2}\right)\alpha_{p,n}\right], \quad p = 0, 1, 2, \ldots, (r-2) \tag{27}$$

$$\sin(t\theta) = 2\sin[(t-1)\theta]\cos\theta - \sin[(t-2)\theta] \tag{28}$$

$$\cos(t\theta) = 2\cos([t-1)\theta]\cos\theta - \cos[(t-2)\theta] \tag{29}$$

By using the relations given in Eqs. (28), (29), we can rewrite Eq. (27) as

$$\begin{aligned} H_{jn}^{P}(k) = [2(-1)^n \cos\alpha_{p,n} \left\{ \sum_{i=0}^{j} w_p(j-i) \sin\left(i - \frac{1}{2}\right)\alpha_{p,n} \right\} \\ - (-1)^n \sum_{i=0}^{j} w_p(j-i)\sin(i-3/2)\alpha_{p,n}], \quad p = 0, 1, 2, \ldots, (r-2) \end{aligned} \tag{30}$$

$$\begin{aligned} H_{jn}^{P}(k) = [2(-1)^n \cos\alpha_{p,n}\{-w_p(j)\sin\frac{\alpha_{p,n}}{2} + \sum_{i=1}^{j} w_p(j-i)\sin(i-1/2)\alpha_{p,n}\} \\ + (-1)^n w_p(j)\sin\left(3\alpha_{p,n}/2\right) + (-1)^n w_p(j-1)\sin\frac{\alpha_{p,n}}{2} \\ - (-1)^n \sum_{i=2}^{j} w_p(j-i)\sin(i-3/2)\alpha_{p,n}\}], \quad p = 0, 1, 2, \ldots, (r-2) \end{aligned} \tag{31}$$

$$\begin{aligned} H_{jn}^{P}(k) = \Bigg[(-1)^n \sin\left(\frac{\alpha_{p,n}}{2}\right)\{w_p(j) + w_p(j-1)\} \\ +2H_{(j-1)n}^{P}(k)\cos\left(\alpha_{p,n}\right) - H_{(j-2)n}^{P}(k)\Bigg], \quad p = 0, 1, 2, \ldots, (r-2) \end{aligned} \tag{32}$$

$$Y(k) = Y(2^p(2n+1)) = H_{jn}^{P}(k) + H_{jn}^{P}(k+1), \quad p = 0, 1, 2, \ldots, (r-2) \tag{33}$$

From Sect. 2, we get 16-point MDCT architecture shown in Fig. 1 which consists of two units connected serially (1) preprocessing unit and (2) IIR filter structure module. The preprocessing unit stated in Eqs. (4), (5), (9b), (22b), (23b) involves $\frac{N}{2}$ multiplications and $\frac{N}{2} + \sum_{m=1}^{m=r-1} \frac{N}{2^m}$ additions to produce input for the IIR filter structure module. Equations (32), (33) are used to generate IIR filter structure of Fig. 2. It comprises of four adders, two delay elements and two multipliers for implementation. Further, it is to be emphasized here that all the coefficients would be computed serially by using the same structure. Since the input sequence is folded in every subsequent group (p), the computation cycles will reduce by the factor of 2 with each successive group. Therefore, the filter structure requires $\frac{N}{2^{p+2}}$ cycles to compute any coefficient of MDCT. For example for $N = 16$, p values will be 0, 1, 2. All odd-indexed coefficients $Q(1), Q(3), Q(5)$ and $Q(7)$ will be computed through IIR filter

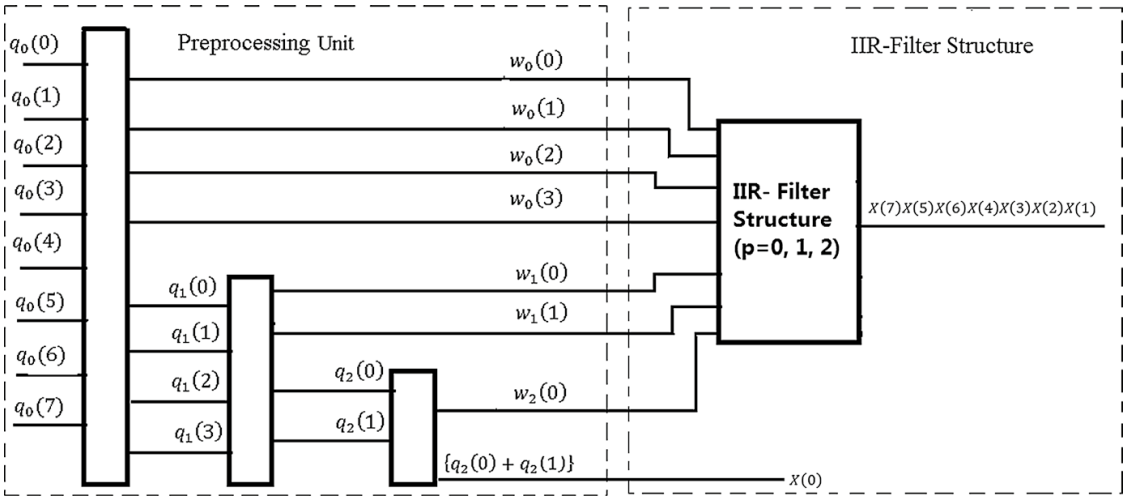
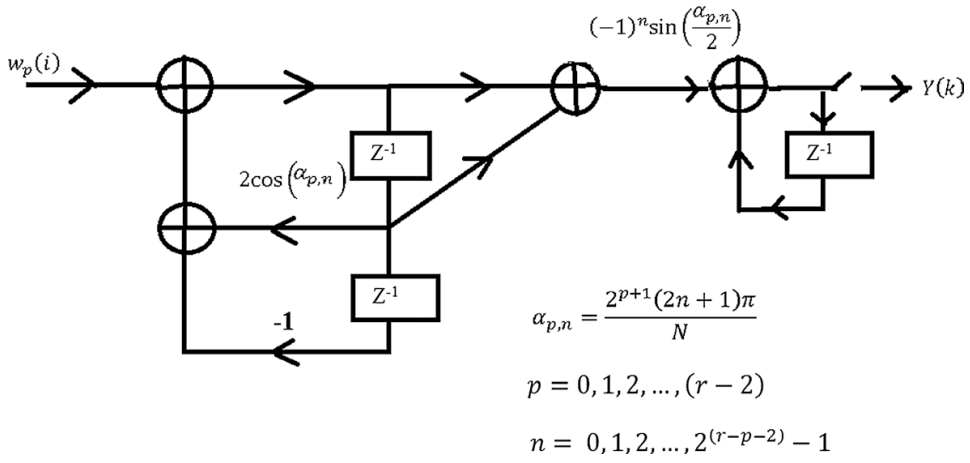


Fig. 1 Recursive architecture for 16-point MDCT



$$\alpha_{p,n} = \frac{2^{p+1}(2n+1)\pi}{N}$$

$$p = 0, 1, 2, \ldots, (r-2)$$

$$n = 0, 1, 2, \ldots, 2^{(r-p-2)} - 1$$

Fig. 2 IIR filter structure of MDCT

structure module with $p = 0$, and for obtaining one odd-indexed coefficient $\frac{N}{4}$ (=four) computation cycles will be needed. For computation of even-indexed coefficients $Q(2)$, and $Q(6)$ which belongs to $p = 1$ group, $\frac{N}{8}$ (=two) computation cycles will be required to compute each coefficient. In the similar manner, for $p = 2$ group, $Q(4)$ coefficient is calculated in $\frac{N}{16}$ (=one) computational cycle. No computational cycle is required for computing $Q(0)$ as it is computed directly through preprocessing stages only. It can be concluded that the maximum computation cycles required by realized IIR filter structure are $\frac{N}{4}$ for a transform coefficient that belongs to $p = 0$ group.

This IIR filter structure can be connected in parallel to increase the data throughput. For N = 16, three such filters can be connected in parallel to increase the data throughput threefold.

Table 1 Hardware comparison to compute MDCT

	[5]	[6]	[12]	[13]	[15]	[16]	[17]	[3, 19–21]	Proposed
Multipliers	5	3	2	3	2	2+1*	7	3	2
Adders	9	3	11	5	3+3*	5	11	3	4
Latches	6	3	6	3	4	4	4	2	3
Computational cycle	$\frac{N}{4}+1$	N	$\frac{N}{4}$	$\frac{N}{4}$	$\frac{N}{2}$	$\frac{N}{8}-\frac{1}{2}$	$\frac{N}{16}+2$	N	$\frac{N}{2^{p+2}}$
DTPT (data throughput)	2	1	2	1	1	2	2	1	1

*Additional adders and a multiplier are required to implement CSD (canonic signed digit) for recursive cosine generation

3 Discussion and Comparison

Performance of the proposed algorithm is compared with the existing recursive algorithms in terms of various performance parameters like hardware components (multipliers, adders and latches) and execution cycles as shown in Table 1. The recursive structure introduced by Chen and Liu [6] requires three multipliers, three adders and three latches for implementation. Nikolajevic and Fettweis implemented MDCT hardware accelerator [19–21] which are one latch deficient, respectively, than [6]. The computation cycles in [5, 12, 13] are one quarter that of computational cycles in [6, 19–21]. Even though the data throughput (DTPT) is twice in [5, 12, 17], the hardware requirement is highest in [5, 12, 17] among all other structures. Hardware complexity of [15, 16] is lower than [5, 6, 12, 17, 19–21], and also the computational cycles are half that of [5, 12] and one eighth that of [6, 19–21]. It is observed from Table 1 that hardware requirement in the proposed structure is least compared to the existing structures [5, 6, 12, 13, 15–17, 19–21]. The number of computational cycles required in the existing structures remains fixed, but it is variable in the proposed structure and decreases with the increase in group number p,, thereby reducing computation time and hardware cost. This is a well-known fact that due to finite word length limitation in VLSI implementation, large number of computational cycles will increase the truncation error which in turn degrades the precision of the results [29]. Therefore, truncation error will be minimum in the proposed structure. Hence, the computation results using the proposed structure will be more accurate in comparison of other structures. In VLSI designing, number of multipliers required to be fabricated on the chip plays a very crucial role in deciding the area–time complexity of the implemented structure as multiplier occupies more chip area as compared to delay or adder operators.

Table 2 illustrates the number of multiplication and addition operations required by available and the proposed recursive structures module for ($N = 2^n$, where $n = 3, 4, \ldots, 9$). Following is the observation of Table 2:

- Structure proposed by Chiang and Liu's [6] requires $\frac{N}{2}(2N+1)$ multiplications and $\frac{N}{2}(3N)$ additions to compute all coefficients of N-point MDCT which are highest compared to all other structures.

Table 2 Number of multiplication/additions in recursive kernel to evaluate N-point MDCT coefficients

Mult./add.	[5]	[6]	[12]	[13]	[15]	[19–21]	Proposed
N	$\frac{N}{4}\left(\frac{N}{2}+2\right)\Big/\frac{N}{4}(N+4)$	$\frac{N}{2}(2N+1)\Big/\frac{N}{2}(3N)$	$\frac{N}{2}\left(\frac{N}{2}+1\right)\Big/\frac{N}{2}\left(\frac{5N}{2}+2\right)$	$\frac{N}{2}\left(\frac{N}{4}+1\right)\Big/\frac{N}{2}\left(\frac{N}{2}+2\right)$	$\frac{N^2}{2}\Big/\frac{N}{2}(3N+2)$	$\frac{N}{2}(N+2)\Big/\frac{N}{2}(2N+1)$	$\sum_{m=0}^{r-2}\frac{N}{2^{m+2}}\left(\frac{N}{2^{m+2}}+1\right)\Big/\left[\sum_{m=0}^{r-2}\left(\frac{N}{2^{m+1}}+2\right)\frac{N}{2^{m+2}}\right]$
8	12/24	68/96	20/88	12/24	32/104	40/68	8/16
16	40/80	264/384	72/336	40/80	128/400	144/264	28/56
32	144/288	1040/1536	272/1312	144/288	512/1568	544/1040	100/200
64	544/1088	4128/6144	1056/5184	544/1088	/5308	2112/4128	372/744
128	2112/4224	16,448/24,576	4160/20,608	2112/4224	8192/24,704	8320/16,448	1428/2856
256	8320/16,640	65,664/98,304	16,512/82,176	8320/16,640	32,768/98,560	33,204/65,664	5588/11,176
512	33,024/66,048	262,400/393,216	65,792/328,192	33,024/66,048	131,072/393,728	131,584/262,400	22,100/44,200

- Computational complexity of [19–21] is same, i.e., $\frac{N}{2}(N+2)$ multiplications and $\frac{N}{2}(2N+1)$ additions.
- Lai's algorithm [12] takes $\frac{N}{2}\left(\frac{N}{2}+1\right)$ multiplication and $\frac{N}{2}\left(\frac{5N}{2}+2\right)$ additions which are lesser than [6] and [19–21] but greater than [5, 13].
- Algorithm [5, 13] requires the same number of multiplication and addition. Among all the existing structures, the number of multiplications and addition is least in [5, 13].
- 512-point MDCT computation in [5, 13] requires 33,024 multiplications and 66,048 additions. The proposed algorithm takes only 22,100 multiplications and 44,200 additions. In comparison with [5, 13], the suggested structure shows a same percentage of reduction in multiplication and addition which is 33.07%.

Fast algorithms for MDCT presented in [25] require less flop counts (total number of additions and multiplications operations) than that of the proposed recursive algorithm. The total number of adders and multipliers for implementation of algorithm [25] is given by $\frac{17}{18}N\log_2\left(\frac{N}{2}\right)+\frac{29}{27}N+\frac{2}{9}(-1)^{\log_2\left(\frac{N}{2}\right)}\log_2\left(\frac{N}{2}\right)-\frac{4}{27}(-1)^{\log_2\left(\frac{N}{2}\right)}$, where N is the input data size, which are more than that required in the proposed structure. It is well known that the structural changes are required to implement fast algorithm of varying input sample sizes, whereas fixed recursive structure can be utilized to the compute results for any length of input sequence. Therefore, hardware complexity is higher for fast algorithms compared to the recursive structures. To elucidate the computation, a comparison is placed in Table 3 for $N=1024$. It can be observed that after adding preprocessing computations, the proposed algorithm requires least flop counts among other well-known recursive algorithms [5, 6, 12, 13, 15, 19–21].

Results of simulation of the proposed algorithm are shown in "Appendix" for $N=16$. Simulated results show that the algorithm works for computation of MDCT coefficients.

4 Conclusions

This paper proposes an algorithm for MDCT and its corresponding structure. The suggested structure requires the least hardware as well as computational cycles compared to any of the previously reported recursive MDCT structures in the literature. The computational cycles required to compute MDCT coefficients are not fixed in the presented algorithm, and it decreases with the increase in the group number. Further with the use of suggested algorithm, computation time and number of multiplication and addition operations are reduced in calculating transform coefficients. Also, any desired coefficient can be computed independently by the suggested structure. The fewer cycles facilitate the implementation of real-time audio codec systems in VLSI. Thus, the structure presented in paper is a better solution for various audio compression applications.

Table 3 Comparison of computational complexity for various recursive MDCT architecture ($N = 1024$)

References	Preprocessing		Recursive structure		Total = Preprocessing + recursive structure		Flop counts = Total multiplications + total additions
	Multiplications	Additions	Multiplications	Additions	Multiplications	Additions	
[5, 13]	512	764	131,584	263,168	132,096	263,932	396,028
[6]	0	0	1,049,088	1,572,864	1,049,088	1,572,864	1,049,088
[19–21]	0	0	525,312	1,049,088	263,168	1,049,088	1,574,400
[12]	0	512	262,656	1,311,744	262,656	1,312,256	1,574,912
[15]	0	512	524,288	1,573,888	524,288	1,574,400	2,098,688
Proposed	512	1534	87,892	110,024	8,8404	111,558	199,962

Acknowledgements The author would like to acknowledge the anonymous reviewers for their valuable comments. This study was supported by University Grant of Commissions, Government of India [Grant No. 3579(NET-DEC.2013)].

Appendix

Computation of MDCT Coefficients for an Input Sequence $x(n)$ Using MATLAB 8.5 Version

Given a sequence of $N = 16$ samples

$$x(n) = [9, 5, 10, 4, 8, 12, 6, 7, 23, 17, 11, 32, 15, 8, 16, 17], \quad n = 0, 1, 2, \ldots, 15 \tag{34}$$

Computation of $X(k)$ using Eq. (1)

$$X(k) = [-125.1546, -39.4699, 15.3834, 6.2849, -46.5619, -35.1487, 1.7654, 15.2230], \quad k = 0, 1, 2, \ldots, 7 \tag{35}$$

Changing the order of input sequence by applying Eqs. (4) and (5) in Eq. (34)

$$g(i) = [-15, -8, -16, -17, 9, 5, 10, 4, 8, 12, 6, 7, 23, 17, 11, 32], \quad i = 0, 1, 2, \ldots, 15 \tag{36}$$

From Eq. (23b), we get

$$w_0(n) = [-3.2091, -6.4826, -17.6485, -27.4492], \quad n = 0, 1, 2, 3 \tag{37}$$

$$w_1(n) = [-19.7217, 6.3975], \quad n = 0, 1 \tag{38}$$

$$w_2(n) = [-35.1727], \quad n = 0 \tag{39}$$

$$x^3(0) = [-101.4571] \tag{40}$$

Using Eq. (32) to obtain $H^p_{ji}(k)$

$$H^p_{3i}(k) = [-23.6976, 31.1557, -21.6911, 15.2230], \quad i = 0, 1, 2, 3 \tag{41}$$

$$H^p_{1i}(k) = [-15.7723, -13.4577], \quad i = 0, 1, 2 \tag{42}$$

$$H^p_{0i}(k) = [-24.8708], \quad i = 0 \tag{43}$$

By Eq. (33), we get MDCT coefficients:

$$X(k) = [-125.1546, -39.4699, 15.3834, 6.2849, -46.561, -35.1487, 1.7654, 15.2230], \quad k = 0, 1, 2, \ldots, 7 \tag{44}$$

From Eq. (44), we obtained the MDCT coefficients as given in (35).

References

1. V. Britanak, H.J.L. Arriëns, Fast computational structures for an efficient implementation of the complete TDAC analysis/synthesis MDCT/MDST filter banks. Signal Process. **89**(7), 1379–1394 (2009)

2. V. Britanak, An efficient computing of oddly stacked MDCT/MDST via evenly stacked MDCT/MDST and vice versa. Signal Process. **85**(7), 1353–1374 (2005)
3. V. Britanak, A survey of efficient MDCT implementations in MP3 audio coding standard: retrospective and state-of-the-art. Signal Process. **91**(4), 624–672 (2011)
4. Y.H. Chan, L.P. Chau, W.C. Siu, Efficient implementation of discrete cosine transforms using recursive filter structure. IEEE Trans. Circuits Syst. Video Technol. **4**(6), 550–552 (1994)
5. C.H. Chen, B.D. Liu, J.F. Yang, Recursive architectures for realizing modified discrete cosine transform and its inverse. IEEE Trans. Circuits Syst. II Analog Digit. Signal Process. **50**(1), 38–45 (2003)
6. H.C. Chiang, J.C. Liu, Regressive implementations for the forward and inverse MDCT in MPEG audio coding. IEEE Signal Process. Lett. **3**(4), 116–118 (1996)
7. D.F. Chiper, An efficient algorithm for a memory-based systolic array VLSI implementation of type IV DCT, in *International Symposium on Signals, Circuits and Systems* (2015), pp. 1–4
8. D.F. Chiper, A parallel VLSI algorithm for a high throughput systolic array VLSI implementation of type IV DCT, in *International Symposium on Signals, Circuits and Systems* (2009), pp. 1–4
9. D.F. Chiper, A new VLSI algorithm for a high-throughput implementation of type IV DCT, in *International Conference on Communications* (2016), pp. 17–20
10. D.F. Chiper, A new VLSI algorithm and architecture for the hardware implementation of type IV discrete cosine transform using a pseudo-band correlation structure. Open Comput Sci **1**(2), 243–250 (2011)
11. R.K. Chivukula, Y.A. Reznik, Efficient implementation of a class of MDCT/IMDCT filters banks for speech and audio coding applications, in *IEEE International Conference on Acoustics, Speech and Signal Processing* (2008), pp. 213–216
12. S.C. Lai, S.F. Lei, C.H. Luo, Common architecture design of novel recursive MDCT and IMDCT algorithms for application to AAC, AAC in DRM, and MP3 codecs. IEEE Trans. Circuits Syst. II Express Briefs **56**(10), 793–797 (2009)
13. S.F. Lei, S.C. Lai, Y.T. Hwang, C.H. Luo, A high-precision algorithm for the forward and inverse MDCT using the unified recursive architecture, in *IEEE International Symposium on Consumer Electronics* (2008), pp. 1–4
14. S.F. Lei, S.C. Lai, P.Y. Cheng, C.H. Luo, Low complexity and fast computation for recursive MDCT and IMDCT algorithms. IEEE Trans. Circuits Syst. II Express Briefs **57**(7), 571–575 (2010)
15. S.C. Lai, Y.P. Yeh, W.C. Tseng, S.F. Lei, Low-cost and high-accuracy design of fast recursive MDCT/MDST/IMDCT/IMDST algorithms and their realization. IEEE Trans. Circuits Syst. II Express Briefs **59**(1), 65–69 (2012)
16. S.C. Lai, Y.P. Yeh, S.F. Lei, Coefficient-free and high-accuracy recursive kernel design for forward and inverse transforms of MDST and MDCT, in *International Conference on Communications, Circuits and Systems* (2013), vol. 1, pp. 38–41
17. L. Li, H. Miao, X. Li, D. Guo, Efficient architectures of MDCT/IMDCT implementation for mpeg audio codec, in *International Conference on Anti-counterfeiting, Security, and Identification in Communication* (2009), pp. 156–159
18. J. Matsumoto, S. Omori, M. Nishiguchi, K. Iijima, Voice encoding method and apparatus using modified discrete cosine transform. U.S. Patent No. 5,819,212 (1998)
19. V. Nikolajevic, G. Fettweis G, New recursive algorithms for the forward and inverse MDCT, in *IEEE Workshop on Signal Processing Systems* (2001), pp. 51–57
20. V. Nikolajevic, G. Fettweis, Computation of forward and inverse MDCT using Clenshaw's recurrence formula. IEEE Trans. Signal Process. **51**(5), 1439–1444 (2003)
21. V. Nikolajevic, G. Fettweis, New recursive algorithms for the unified forward and inverse MDCT/MDST. J. VLSI Signal Process. **34**(3), 203–208 (2003)
22. S.M. Perera, Signal flow graph approach to efficient and forward stable DST algorithms. Linear Algebra Appl. **542**, 360–390 (2018)
23. S.M. Perera, A. Madanayake, N. Dornback, N. Udayanga, Design and digital implementation of fast and recursive DCT II–IV algorithms. Circuits Syst Signal Process. **38**(2), 529–555 (2019)
24. J.P. Princen, M.H. Chan, System and method for noise reduction using bi-orthogonal modified discrete cosine transform. U.S. Patent No. 6,636,830 (2003)
25. X. Shao, S.G. Johnson, Type-IV DCT, DST, and MDCT algorithms with reduced numbers of arithmetic operations. Signal Process. **88**(6), 1313–1326 (2008)
26. H. Shu, X. Bao, C. Toumoulin, L. Luo, Radix-3 algorithm for the fast computation of forward and inverse MDCT. IEEE Signal Process. Lett. **14**(2), 93–96 (2007)

27. T.K. Truong, P.D. Chen, T.C. Cheng, Fast algorithm for computing the forward and inverse MDCT in MPEG audio coding. Signal Process. **86**(5), 1055–1060 (2006)
28. C.T. Wang, T.S. Chen, W.H. Chao, A new audio watermarking based on modified discrete cosine transform of MPEG/audio layer III, in *IEEE International Conference on Networking, Sensing and Control* (2004), vol. 2, pp. 984–989
29. J.L. Wang, C.B. Wu, B.D. Liu, J.F. Yang, Implementation of the discrete cosine transform and its inverse by recursive structures, in *IEEE Workshop on Signal Processing Systems* (1999), pp. 120–130
30. J.S. Wu, H.Z. Shu, L. Senhadji, L.M. Luo, A fast algorithm for the computation of 2-D forward and inverse MDCT. Signal Process. **88**(6), 1436–1446 (2008)

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

JOURNAL OF Applied Polymer SCIENCE

Flame retardancy of ceramic-based rigid polyurethane foam composites

Anuja Agrawal,[1] Raminder Kaur ,[1] Ravinderjit Singh Walia[2]
[1]Department of Applied Chemistry & Polymer Technology, Delhi Technological University, Delhi 110042, India
[2]Department of Mechanical Engineering, Delhi Technological University, Delhi-110042, India
Correspondence to: R. Kaur (E-mail: rkaurdtu@gmail.com)

ABSTRACT: In this work, ceramic fillers zirconia and alumina powder were incorporated in the rigid polyurethane foams derived from modified castor oil and their impact on the mechanical, thermal, and fire performances of composite foams have been analyzed. It was observed that the addition of ceramic filler showed improved mechanical and thermal properties and best properties were shown by 6% zirconia with compressive strength of 6.61 MPa and flexural strength of 5.72 MPa. Zirconia also demonstrated an increase in $T_{5\%}$ up to 260 °C. Cone calorimetry shows a decrease in peak of heat release from 118 to 84 kW m^{-2} and 94 kW m^{-2} by the incorporation of alumina and zirconia powder, respectively. Furthermore, total heat release (THR), smoke production rate (SPR), and total smoke release (TSR) were also found to decrease remarkably on the incorporation of ceramic fillers. So, these fillers have a great potential as an additive to incorporate good mechanical, thermal, and fire properties in bio-based rigid PU foams. © 2019 Wiley Periodicals, Inc. J. Appl. Polym. Sci. **2019**, *136*, 48250.

KEYWORDS: biomaterials; flame retardance; foams; polyurethane; thermal properties

Received 26 November 2018; accepted 20 June 2019
DOI: 10.1002/app.48250

INTRODUCTION

Rigid polyurethane foam (RPUF) has a wide range of applications in disposable packaging, thermal insulation, construction and refrigeration industry, prototypes, and industrial patterns due to its some unique properties such as low density, thermal conductivity, water absorption and comparatively high compressive strength.[1] A urethane linkage is generated by the reaction of hydroxyl with the isocyanate, which is the basic unit of polyurethane structure. However, due to diminishing resources, petroleum-based polyols are being replaced by the renewable raw materials.[2–4] Utilization of renewable sources in place of petroleum-based raw material and incorporation of blowing agent makes them more flammable.[5] However, the application of RPUF requires low flammability, so several studies have been conducted for abating flammability by introducing some flame retardant groups in the foam structure or by incorporating some unique flame retardants.[6–8] Zhang *et al.*[9] prepared a flame retarded polyol castor oil phosphate polyol (COFPL) by the introduction of phosphate group in the polyol derived from castor oil and developed flame-retarded RPUF using COFPL. In addition, different types of flame retardants such as chlorinated flame retardants, brominated flame retardants, phosphorus-based compounds, metal hydroxide, nitrogen-based compounds, nanomaterials, or boron compounds are found to be a promising candidate to reduce the flammability. Char-forming property of RPUF was also enhanced by the incorporation of phosphorus flame retardant in the foam.[10] Results concluded that the char yield and the thermal stability of foams was improved by the incorporation of the flame retardant. Furthermore, an additional nanoclay filler, organically modified montmorillonite, was employed to further enhance the fire properties of foams. The prepared foams exhibited reduction in heat release rate (HRR) and total smoke production (TSP), while an increase in combustion duration time was observed. Hu *et al.*[11] analyzed the effect of incorporation of dimethyl methylphosphonate (DMMP) and expandable graphite (EG) in PU foam as flame retardants. A significant improvement in the flame retardancy of RPUF was observed on the incorporation of DMMP as well as EG. Bian *et al.*[12] enhanced the flame retardancy of RPUF composites by incorporating whisker silicon oxide (WSi) and EG. It was reported that the dynamic mechanical properties as well as flame retardant properties were enhanced on the incorporation of WSi and EG. In a study, Zhang *et al.*[13] synthesized a phosphorus–nitrogen-based flame retardant, namely, DOPO-BA (a condensation product of benzaldehyde, aniline and 9,10-dihydro-9-oxa-10-phosphaphenanthrene-10-oxide) and incorporated it in a rosin-based RPUFs. Results showed that the incorporation of DOPO-BA improved the physical, thermal, and flame retardant properties of RPUF. In a similar study, Yang *et al.*[14] employed hexachlorocyclotriphosphazene, *p*-hydroxybenzaldehyde, and diethyl

© 2019 Wiley Periodicals, Inc.

phosphite to synthesize hexa(phosphitehydroxylmethylphenoxyl) cyclotriphosphazene (HPHPCP). HPHPCP was incorporated into RPUFs prepared by the reaction of polyester polyol and polyether polyol with polymethylene polyphenyl isocyanate. Incorporation of the HPHPCP into the RPUF increased the percentage limited oxygen index (LOI), while the HRR was also found to decrease continuously on increasing the concentration of HPHPCP. Granular silica aerogel has also been used to enhance the flame retardancy of RPUF prepared by employing organic/inorganic polyisocyanurate. It was observed that increasing the concentration of aerogel from 0 to 8 wt % increased the percentage LOI from 29.4 to 34.6%.[15] 3,3′4,4′-biphenyltetracarboxylic dianhydride and 9,10-dihydro-9-oxa-(10-glycidoxypropylene)-10-phosph-phenanthrene-10-oxide were incorporated in RPUF for enhancing the flame retardancy of the foam. It was concluded from the results that the incorporated flame retardant showed a small improvement in the thermal properties but the peak of heat release (PHRR) and total heat release (THR) decreased by approximately 20%.[16] Along with the incorporation of flame retardant, some functional filler such as aluminium diethyl phophinate,[17] calcium hypophosphite,[18] and others may also be used to abate the flammability of RPUF. By selecting appropriate filler, mechanical, thermal, and fire properties of RPUF may be tailored to requisite conditions. However, different types of vegetable oil have been used as polyol for polyurethane applications, castor oil is the most promising as it is the only vegetable oil, which contains hydroxyl groups in its structure naturally, that are essential for generating linkage in polyurethane. Although recent studies had been reported on the castor oil-based polyurethanes reinforced with different flame retardants and has shown enhanced flame retardancy, but still the literature is devoid of much data on castor oil-based RPUF composites reinforced with ceramic fillers such as alumina and zirconia powder. Alumina powder has many unique properties such as hardness, resistance to strong acid and alkali attack at elevated temperatures, excellent dielectric properties, wear resistant, high strength and stiffness, and so forth. High purity alumina is suitable for use up to the temperature of 1925 °C.[19] Zirconia (ZrO_2, zirconium dioxide) is chemically inactive and acid resistant filler. Zirconia is being used as a thermal barrier coating in gas turbine blades, diesel engine parts owing to its low thermal conductivity and in dentistry to prepare dental crown due to its high strength.[20] Owing to these remarkable properties as well as low cost, these fillers are considered as potential engineering ceramics material for properties enhancement.[21]

Verdolotti et al.[22] incorporated alumina powder in flexible PU foam to enhance its hydrophobicity and thermal stability. Swain et al.[23] investigated the influence of nano-silica and nano-alumina on the properties of PU composites. It was observed that both the nanoparticles demonstrated improvement in tensile as well as thermal properties of the composite up to a fixed concentration, then both mechanical as well as thermal properties start to deteriorate due to the agglomeration of the particles. A few researchers employed ZrO_2 to enhance mechanical and thermal properties of PU coating due to the advantageous properties of ZrO_2 such as excellent thermal stability, chemical inertness and high hardness.[24–26] Hybrid of alpha-zirconium phosphate and nano-graphene oxide has also been incorporated in the rigid polyvinyl chloride foam to enhance the flame retardancy of the foam.[27] So, alumina and zirconia powder may prove a better additive to enhance the mechanical and thermal properties of RPUF. But still, no study has been conducted on RPUF incorporated with ceramic fillers to enhance the flame retardancy of RPUF. The present study is an attempt to overcome such research gaps with an aim to produce RPUF composite with enhanced mechanical and structural properties using these ceramic fillers. Additionally, the effect of variation in concentration of zirconia and alumina powder on thermal, mechanical, and fire performances of RPUF has also been analyzed.

EXPERIMENTAL

Materials

Castor oil (99.9%) was purchased from Thomas Baker Chemicals Pvt., Ltd. (Mumbai, India) and 4,4′-diphenylmethane diisocyanate (AR grade) was obtained from Krishna Enterprises (Delhi, India). *n*-Pentane (99.5%) and silicon oil were purchased from RANKEM, Gurugram, India. Glycerol (99%) was supplied by Fischer Scientific, Mumbai, India. Catalyst DABCO 33 LV was purchased from Sigma-Aldrich, Delhi, India. Alumina (50–200 μm) was obtained from Fisher Scientific, Mumbai, India and zirconia powder (in the range of 3 μm, with ZrO_2: ≥98%) was purchased from Zirox Technologies (Bewar, India). All the reagents were of analytical grade and used as supplied.

Methods

Foam Preparation. The RPUF composites were obtained by the same method as reported in the literature,[28,29] in brief, modified polyol is prepared by the transesterification of castor oil under the atmosphere of nitrogen using 2:1 ratio of castor oil to glycerol at the temperature 180–200 °C for 4 h to attain the hydroxyl value of 350–450 mg KOH equiv g^{-1}. The number average molecular weight of the modified polyol was determined using gel permeation chromatography and recorded as 1165 g/mol. The predetermined quantity of the filler was added to the modified polyol taken in a paper cup. Prior to addition, fillers were surface treated with 3-aminopropyltriethoxysilane (APTES) and calculated amount of other ingredients was added to the contents of the cup and thoroughly mixed to form polyol premix. A calculated amount of MDI was then added to the cup. The resulted reaction mixture was poured into an aluminum mold coated with releasing agent, that is, silicon oil. To ensure complete curing, the molds were left at room temperature for 72 h. Developed foams were cut into desired dimensions and tested for their mechanical, thermal, and flame retardant properties. Table I shows the formulation of castor oil-based

Table I. Formulation of RPUF Modified with Fillers

Material	Amount
Modified polyol	20 g
DABCO 33 LV	2 pbw
n-Pentane (blowing agent)	10–15 pbw
Silicone oil (surfactant)	3 pbw
Filler	2–10 pbw
MDI	25 g

pbw, parts by weight of polyol.

RPUF composites modified with different concentrations of ceramic fillers.

CHARACTERIZATION

Mechanical Testing

The mechanical properties of the castor oil-based rigid polyurethane samples were determined according to the standard procedures. Testing was conducted on three specimens of each concentration and the average value was reported. Flexural and compressive properties of the resulted foams were measured at room temperature using Instron (model no. 3369) universal testing machine (UTM). Flexural tests were performed according to the ASTM D-790 using specimen with dimensions $80 \times 10 \times 4$ mm^3. The compression tests were performed according to the ASTM D-1621. Specimens of dimensions $25 \times 25 \times 25$ mm^3 were cut from foam in the in-plain direction and tested for 10% compression. The rate of crosshead movement was fixed at 5 mm min^{-1} for each sample.

Thermogravimetric Analysis

Thermogravimetric analysis (TGA) of RPUF incorporated with ceramic fillers was performed employing a thermogravimetric analyzer (PerkinElmer 4000) in nitrogen atmosphere. The heating rate was kept 10 °C min^{-1} and temperature ranging from 50 to 700 °C.

Cone Calorimeter Testing

The antiflammability performance of RPUF samples was analyzed with a cone calorimeter (Jupiter Electronics, Mumbai, India) according to ISO 5660-1 standard at an incident heat flux of 35 kW m^{-2}. The size of samples was $100 \times 100 \times 20$ mm^3. The instrument provided the record of the time to ignition (TTI), total heat release (THR), HRR, smoke production rate (SPR), TSP, and so forth.

RESULTS AND DISCUSSION

RPUF is prepared using polyol derived from castor oil and the ceramic fillers were added to achieve requisite strength and thermal properties. The foam performance depends on its structure, which in turn is affected by the particle size, dispersion, and the properties of the filler incorporated in the reaction mixture. Ceramic fillers can impart stiffness to the material, consequently enhancing the mechanical strength of the rigid foam.[30,31] Effect of incorporation of these fillers on the properties of RPUF is discussed in the subsequent section.

Mechanical Testing

Figure 1 and 2 show the variation in mechanical properties and specific strength (strength/density) on the incorporation of ceramic fillers. Upon reinforcing RPUF with 10% w/w of alumina and 6% w/w of zirconia powder, flexural strength reached up to 2.18 and 5.72 MPa, respectively, as compared to 0.79 MPa of

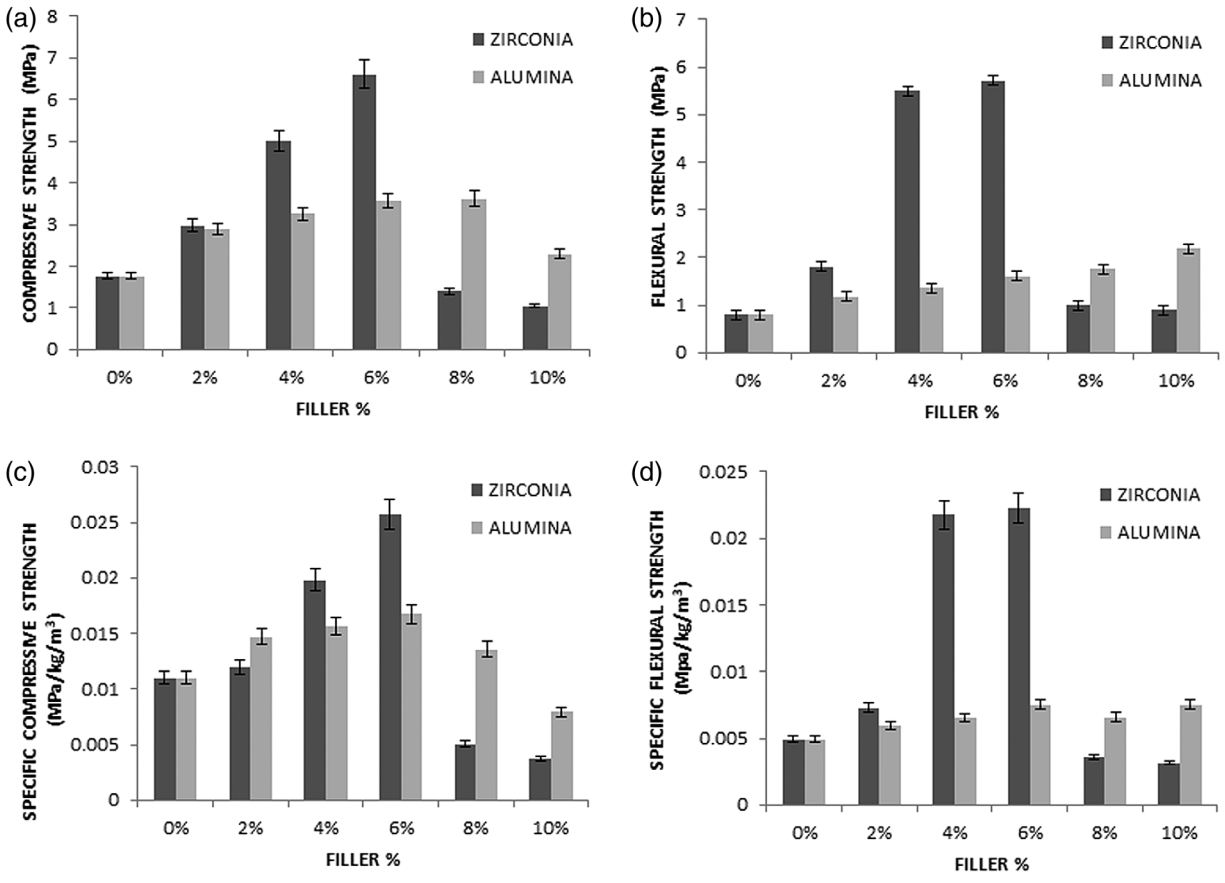


Figure 1. The effect of zirconia and alumina on (a) compressive strength, (b) Flexural strength, (c) Specific compressive strength and (d) Specific flexural strength

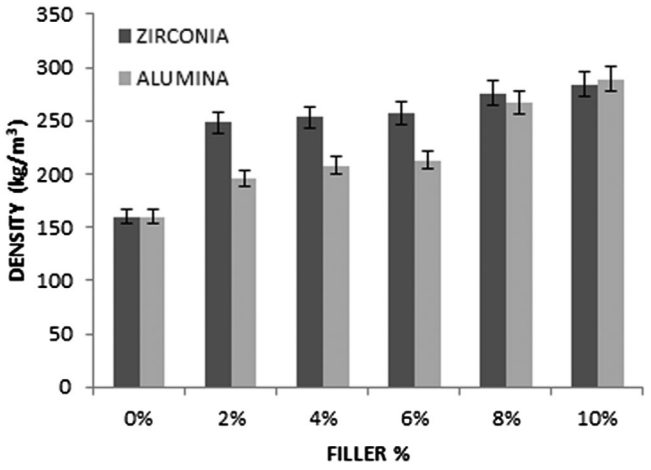
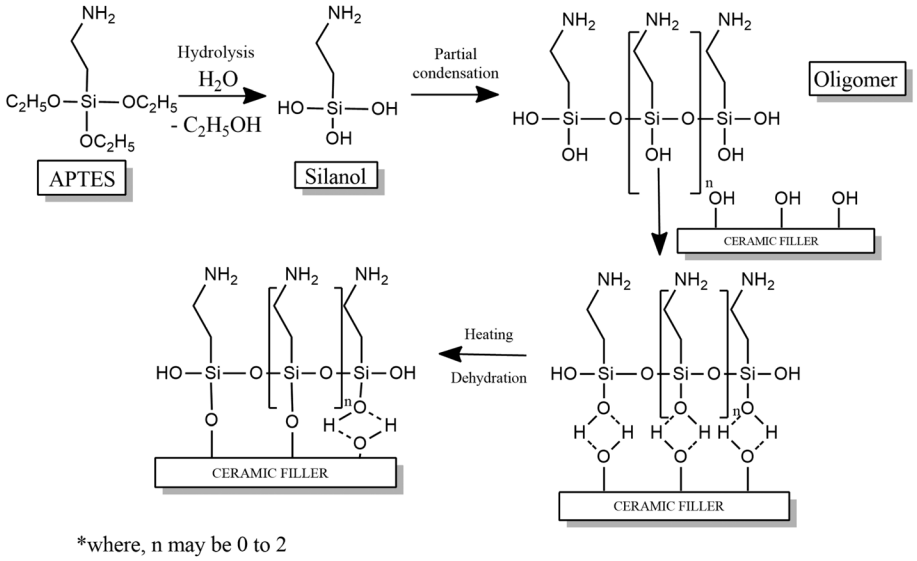


Figure 2. Variation in density with increase in concentration of ceramic fillers.

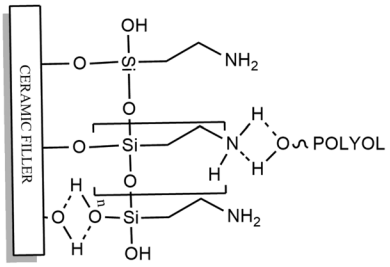
unreinforced polyurethane foam. So the percentage increase in flexural strength was 176 and 624% respectively. Alumina and zirconia powder reinforced PU foam also showed an increase in compressive strength up to 3.62 and 6.61 MPa, respectively, as compare to 1.77 MPa for unreinforced polyurethane foam before the failure occurs, which was, respectively, 104 and 273% higher than that of unreinforced foam.

High-density urethane cellular foam shows little deflection on applying heavy structural loads during construction applications. It also has applications in hand-carved models, prototypes, CNC-machined topographical maps, and industrial patterns.[32] It is observed that the addition of ceramic filler increases the viscosity of the polyol, which indicates more interfacial interaction of the filler with the polyol. APTES treated ceramic fillers have better interaction with the polyol due to the presence of a terminal amino group. APTES can be chemically bonded to the surfaces of the filler by the condensation of its alkoxyl groups with the surface hydroxyl groups of the ceramic filler.[33] (Figure 3). So, increasing the concentration of the zirconia powder up to 6% w/w and alumina powder up to 8% w/w increases the interfacial interaction, consequently enhancing the mechanical properties of RPUF. Beyond these concentrations, particles get agglomerated with each other which deteriorates the mechanical properties of RPUF.[23] It is also observed that the incorporation of zirconia



*where, n may be 0 to 2

(a) Proposed reaction scheme for surface treatment of ceramic filler



(b) Adhesion of surface treated ceramic filler with polyol via intermolecular interaction

Figure 3. (a) Proposed reaction scheme for surface treatment of ceramic filler (b) Adhesion of surface treated ceramic filler with polyol via intermolecular interaction.

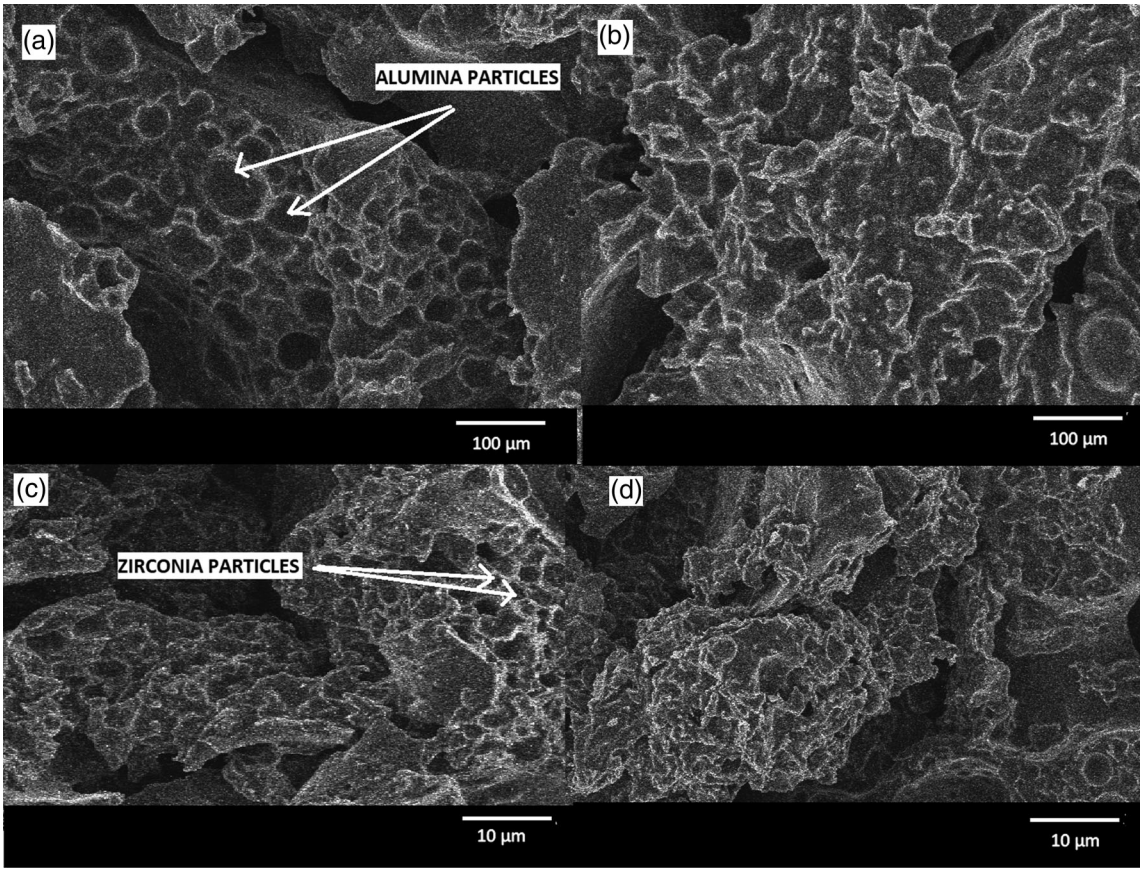


Figure 4. SEM images of RPUF with (a) 8% w/w alumina, (b) 10% w/w alumina, (c) 6% w/w zirconia, and (d) 10% w/w zirconia.

powder shows a much higher increase in the compressive strength as compared to that of alumina powder. This behaviour is attributed to the finer particle size and larger surface area of zirconia powder, which leads to better interaction between the zirconia powder nd polyol. Additionally, alumina powder does not get agglomerated at 8% w/w concentration due to its lower surface area, which promotes better dispersion at a higher concentration in comparison to zirconia powder, consequently enhancing the mechanical properties. Figure 4 shows the dispersion of ceramic filler at different concentrations.

Scanning Electron Microscopy

Figure 5 shows the variation in the cellular morphology of the unreinforced and ceramic filler reinforced RPUF composites. It is evident from the figure that the incorporation of ceramic fillers decreased the cell size from 350 ± 50 μm to 280 ± 20 and 320 ± 20 μm for zirconia and alumina powder, respectively. Cell structure was also more homogeneous and round in shape after the addition of ceramic fillers. It is attributed to the fact that the incorporation of ceramic filler provides more nucleation sites for bubble formation. It is also observed that the ceramic filler increases the viscosity of the polyol, which prevents the bubble from being coalesced and consequently reducing the cell size.

Thermogravimetric Analysis

To evaluate thermal stability of these foam composites, TGA is conducted under the flow of nitrogen. Results showed enhanced thermal properties of RPUFs on the incorporation of alumina

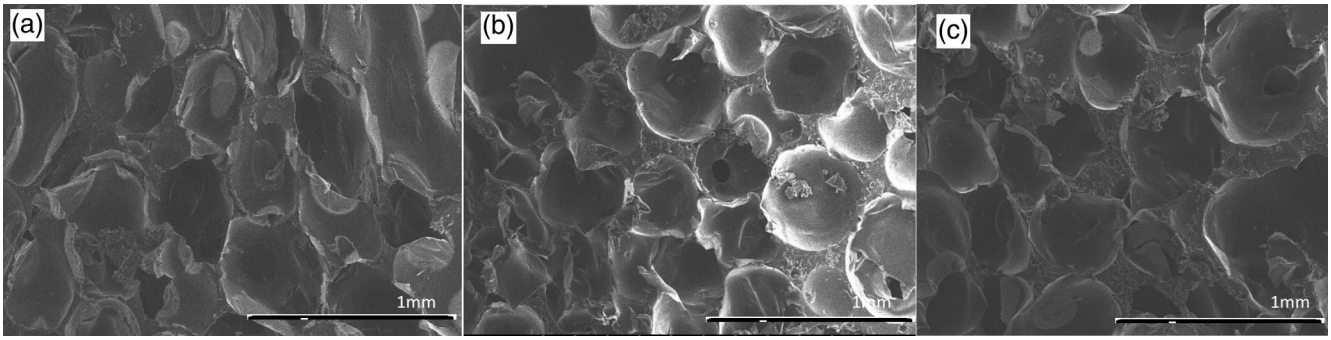


Figure 5. SEM images of RPUF with (a) 0% filler, (b) 6% w/w zirconia and 8% w/w alumina.

and zirconia fillers. Maximum char residue was observed for the alumina at 6% w/w and zirconia at 8% w/w concentration. Usually, the thermal stability of RPUFs is characterized by the temperatures of 5% weight loss ($T_{5\%}$) which is considered as the temperature for onset of degradation.[13] It was noticed that 5% weight loss temperature increased from 192 to 251 °C and 265 °C by the incorporation of alumina and zirconia powder respectively, while there is almost no change in the peaks of TMAX1 and TMAX2. The TGA and differential thermogravimetry (DTG) analysis of the ceramic fillers incorporated RPUF composites are illustrated in Figure 6 and the parameters are summarized in Table II.

It was observed that RPUF shows two weight loss stages, the first peak at around 300–350 °C is related to the degradation of the flexible phase and the decomposition of the urethane linkages in the hard segments. The second peak observed at 450–500 °C is attributed to the degradation of soft segments and the thermolysis of the organic residues. Improvement in thermal stability is associated with strong interfacial interaction between the polyol and the ceramic filler phases. If the surface area of the filler is high, a high polyol-filler interaction restricts the polymer chain mobility, which increases the glass transition temperature, consequently increasing the temperature of 5% weight loss (T5%). Additionally, zirconia powder can act as a cross-linker between PU backbones, consequently enhancing the thermal properties of the foam.[27] The improved thermal stability could also be attributed to the homogeneous dispersion of ZrO_2 within the polyol which acts as a thermal insulator due to its low thermal conductivity.[25]

Cone Calorimeter Testing

Figures 7 and 8 show the cone calorimeter performance of RPUF composites containing varied concentration of ceramic fillers and the results are compiled in Table III. The cone calorimeter test provides quantitative analysis of the flame retardancy and smoke release of foams. TTI is an important fire behavior parameter to judge the flammability of polymeric foams. It is evident that the TTIs of ceramic filler incorporated RPUF composites have shown a considerable increase in comparison with neat polyurethane foam. Incorporation of filler increased TTI from 2 to 6 s for both the foams incorporated with 6% w/w alumina and 6% w/w zirconia powder, which may be attributed to the high temperature

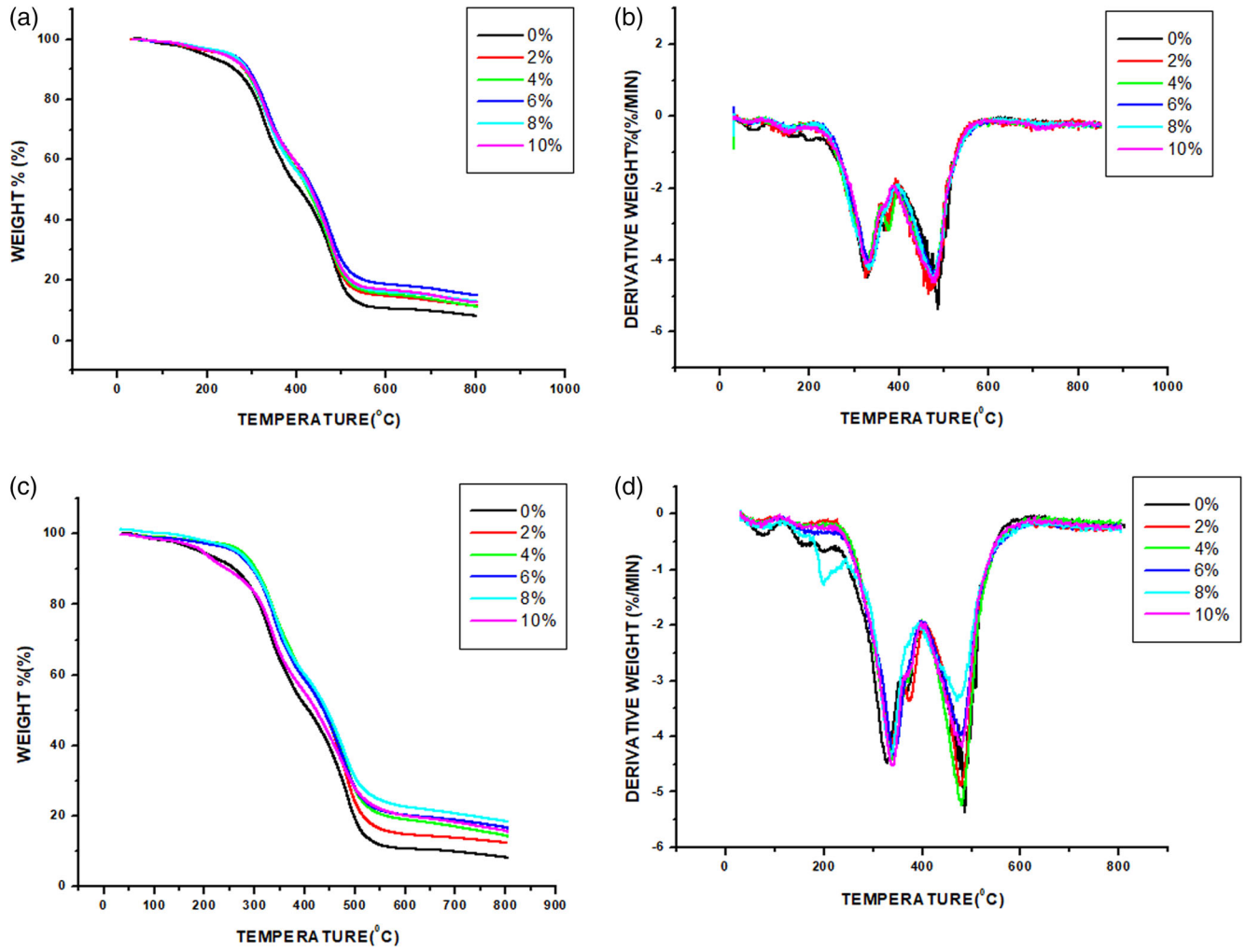


Figure 6. (a) TGA curve of RPUF incorporated with alumina powder, (b) DTG curve of RPUF reinforced with alumina powder, (c) TGA curve of RPUF incorporated with zirconia powder, and (d) DTG curve of RPUF reinforced with zirconia powder. [Color figure can be viewed at wileyonlinelibrary.com]

Table II. Thermal Characteristics of RPUF Incorporated with Ceramic Fillers

Concentration of filler	$T_{5\%}$ (°C)	T_{max1} (°C)	T_{max2} (°C)	Residual mass (%)
0%	192 ± 2	330 ± 4	488 ± 4	8.1 ± 2.0
2% w/w alumina	235 ± 1	326 ± 2	465 ± 2	11.3 ± 1.0
4% w/w alumina	246 ± 1	325 ± 2	474 ± 2	11.2 ± 1.0
6% w/w alumina	251 ± 1	336 ± 1	474 ± 1	14.9 ± 2.0
8% w/w alumina	251 ± 2	334 ± 1	475 ± 1	12.8 ± 1.0
10% w/w alumina	233 ± 2	325 ± 2	480 ± 2	12.5 ± 2.0
2% w/w zirconia	269 ± 2	338 ± 2	478 ± 2	12.5 ± 1.0
4% w/w zirconia	273 ± 2	339 ± 2	480 ± 1	14.3 ± 1.0
6% w/w zirconia	260 ± 2	344 ± 1	477 ± 1	16.6 ± 2.0
8% w/w zirconia	265 ± 2	337 ± 2	471 ± 2	18.4 ± 1.0
10% w/w zirconia	196 ± 2	340 ± 2	481 ± 2	15.7 ± 2.0

performance of these fillers. Conventionally, intensity of fire is correlated with the HRR, Figures 7(a) and 8(a) show the trends in change in HRR with time. It exhibits two peaks for most of the formulations, which are attributed to the formation of a protective charred layer during the burning of foam.[34] Results show that peak heat-release rate (PHRR) was 84 and 94 kW m^{-2} for the foam composite containing 6% w/w of alumina powder and 6% w/w of zirconia powder, respectively, which were less than that of neat RPUF (118 kW m^{-2}). PHRR signifies the surface pyrolysis of RPUF and evolution of a large quantity of flammable

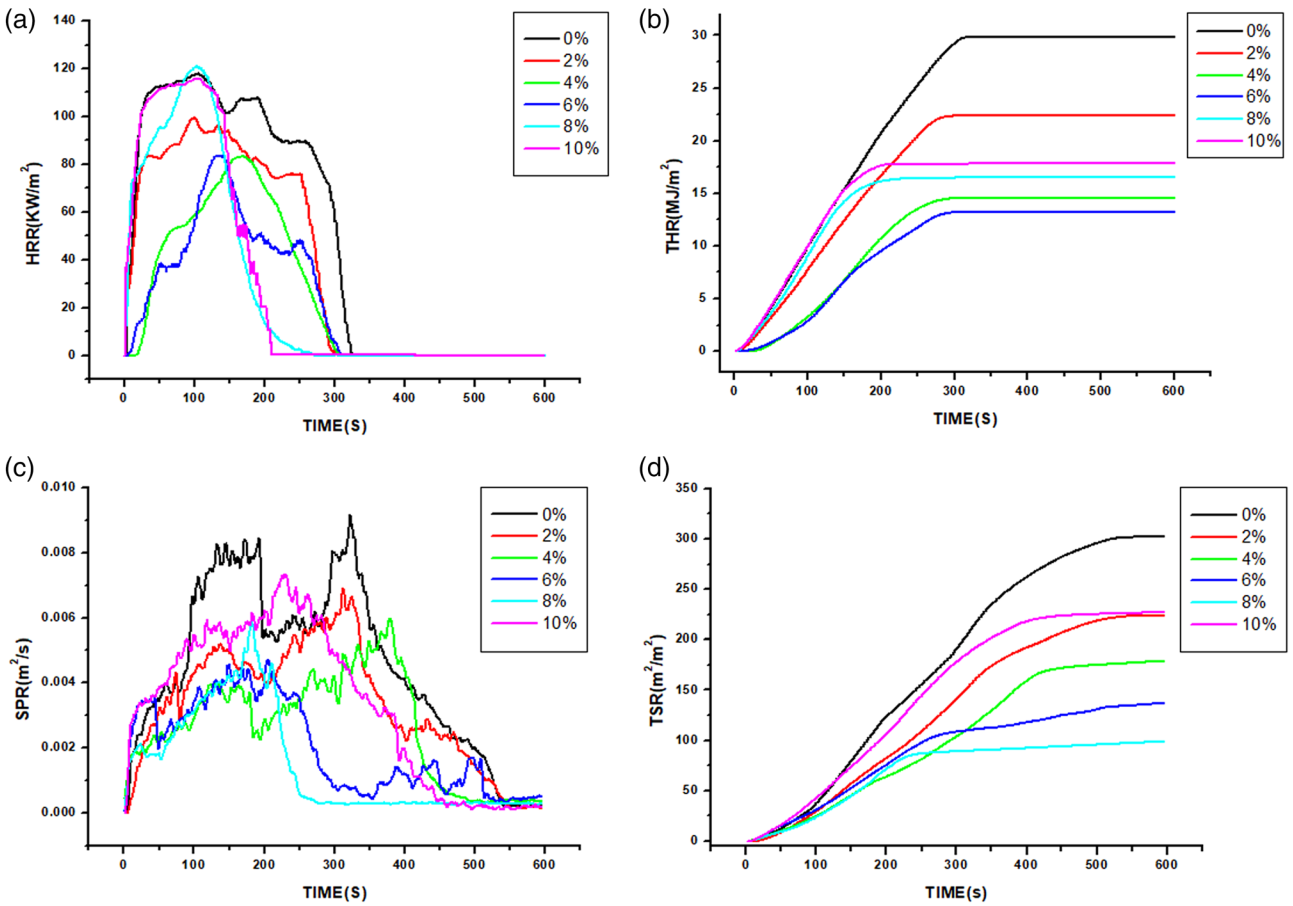


Figure 7. (a) HRR (Heat release rate), (b) THR (Total heat release, (c) SPR (Smoke production rate) and (d) TSR (Total smoke release) of RPUFs incorporated with alumina powder. [Color figure can be viewed at wileyonlinelibrary.com]

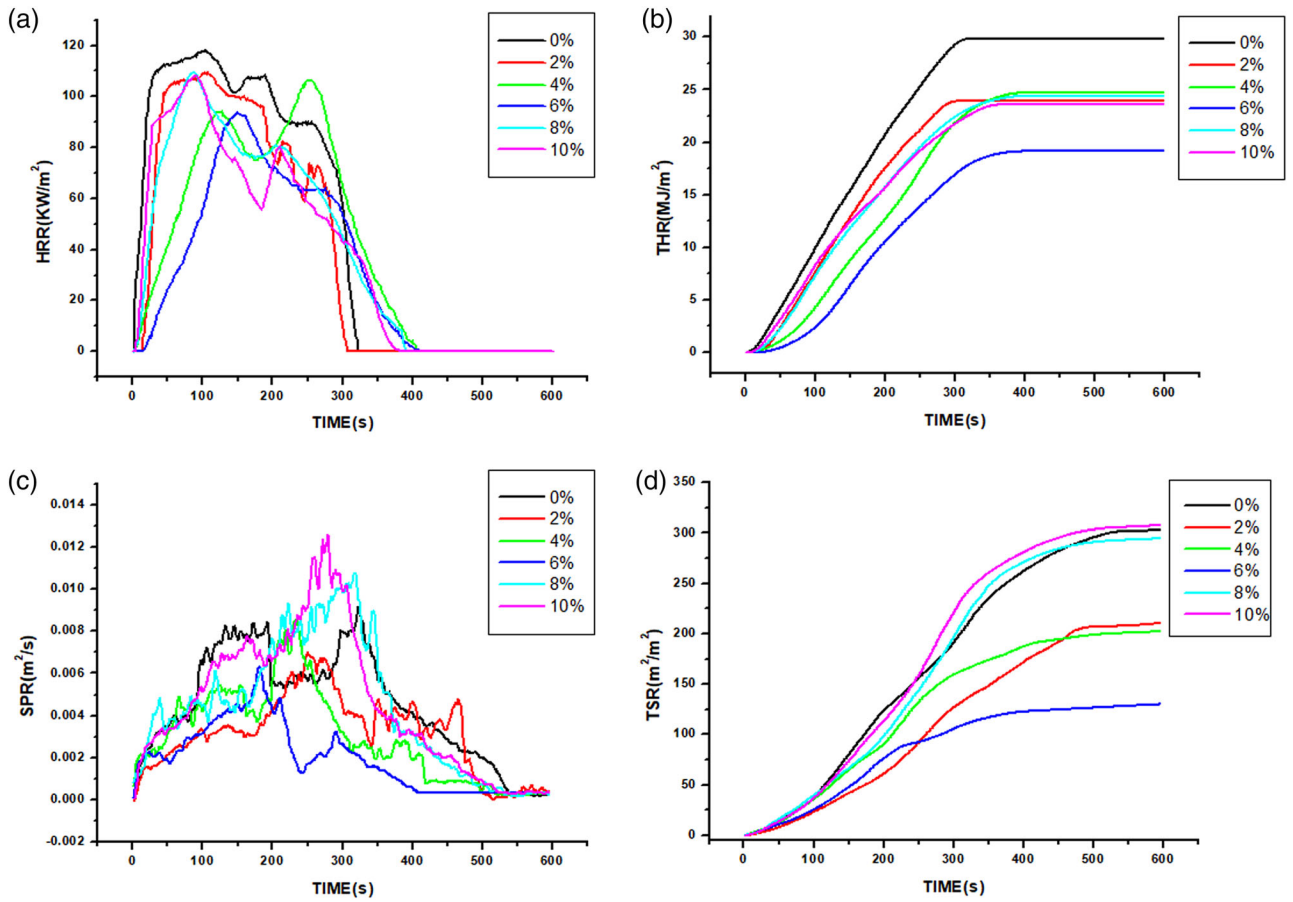


Figure 8. (a) HRR (Heat release rate, (b) THR (Total heat release), (c) SPR (Smoke production rate) and (d) TSR (Total smoke release) of RPUFs incorporated with zirconia powder. [Color figure can be viewed at wileyonlinelibrary.com]

low molecular weight by products such as primary or secondary amine, isocyanate, alcohol, and olefin.[35] Ceramic fillers are high temperature performance fillers and demonstrate better thermal insulation at elevated temperature, which reduces the rate of heat release. Furthermore, on increasing the concentration of filler, agglomeration of the particles deteriorates the insulating effect. Additionally, the APTES treated fillers restrain the segmental motion of the PU backbone owing to increased crosslink density in the RPUF.[33]

Figures 7(b) and 8(b) show that THR also decreased on the incorporation of ceramic fillers. This result illustrates that ceramic powder hinders the heat transfer and flame spread, therefore decreasing the intensity of fire. As the flame retardancy of an additive is also reliant on the augmented char residue

Table III. Fire Characteristics of RPUF Incorporated with Ceramic Fillers

Concentration of filler	PHRR (kW m^{-2})	THR (MJ m^{-2})	TSP (m^2 m^{-2})	Residue (%)	TTI (s)
0%	118 ± 2	29.8 ± 1.0	302 ± 3	11.5 ± 2.0	2 ± 1
2% w/w alumina	100 ± 3	22.4 ± 2.0	224 ± 2	16.4 ± 2.0	4 ± 1
4% w/w alumina	83 ± 2	14.5 ± 1.0	179 ± 3	18.3 ± 1.0	5 ± 1
6% w/w alumina	84 ± 2	13.2 ± 1.0	137 ± 3	20.4 ± 1.0	6 ± 1
8% w/w alumina	121 ± 3	16.5 ± 1.0	99 ± 2	21.4 ± 2.0	5 ± 1
10% w/w alumina	116 ± 2	17.8 ± 2.0	227 ± 2	23.1 ± 1.0	4 ± 1
2% w/w zirconia	109 ± 2	23.9 ± 2.0	210 ± 2	18.9 ± 2.0	3 ± 1
4% w/w zirconia	106 ± 2	24.7 ± 1.0	202 ± 3	19.1 ± 1.0	5 ± 1
6% w/w zirconia	94 ± 2	19.2 ± 1.0	130 ± 2	21.6 ± 2.0	6 ± 1
8% w/w zirconia	109 ± 3	24.4 ± 1.0	294 ± 2	22.1 ± 1.0	6 ± 1
10% w/w zirconia	108 ± 3	23.6 ± 2.0	307 ± 3	24.7 ± 1.0	5 ± 1

yield,[36] the char residue increased from 11.47% for neat RPUF to 23.10 and 24.69% for RPUF with 10% w/w of alumina powder and 10% w/w of zirconia powder, respectively.

The smoke production behavior can be predicted by total smoke release (TSR) and smoke production rate (SPR). Figures 7(c) and 8(c) show the SPR and first peak of SPR denotes that the smoke production decreased from 0.008 to 0.004 m^2 s^{-1} and 0.006 for 6% w/w of alumina powder and 6% w/w of zirconia powder, respectively. The TSR also decreased from 302 to 99 m^2 m^{-2} and 130 m^2 m^{-2} for 8% w/w of alumina powder and 6% w/w of zirconia powder, respectively [Figures 7(d) and 8(d)]. This smoke suppression is attributed to the protective charred layer that prevented the release of the smoke.[37] Increasing the concentration of alumina powder beyond 6% w/w shows an increase in the rate of heat release due to the less insulation caused by the slight agglomeration of alumina powder, whereas smoke release decreases owing to the increased cross-linking density of RPUF. Furthermore, at 10 % w/w concentration of the filler total collapse in the structure demonstrates deteriorated flame retardant properties.

CONCLUSIONS

Rigid PU foams are prepared by incorporating alumina and zirconia powders to enhance the flame retardant and thermal properties of foams. Compressive strength and flexural strength of ceramic filler incorporated foam is found to increase by 273 and 624%, respectively. TGA and DTG curves show increase in thermal stability of ceramic filler reinforced foams. Then, 5% weight loss temperature increases from 192 to 265 °C for the 8% w/w of zirconia powder filled RPUF. TTI is also found to increase from 2 to 6 s for the foam with 6% w/w concentration of zirconia powder as well as alumina powder. Flame retardancy of ceramic filler incorporated RPUFs was also enhanced and 6% w/w concentration of ceramic fillers demonstrated the best properties with up to 29% decrease in HRR and 71% suppression in the production of smoke. As the addition of ceramic powder improves the physical, thermal, and fire properties of vegetable oil-based rigid PU foams, ceramic powder may be considered as a potential additive to improve these properties of rigid PU foam.

ACKNOWLEDGMENT

This work was supported by Defense Research and Development Organization (DRDO), India (grant number: CC/TM/ERIPR/GIA/15-16/029).

REFERENCES

1. Sivertsen, K. 3.063 Polymer Physics; Spring: USA, **2007**.
2. Xue, B.; Wen, J.; Sun, R. *ACS Sustain. Chem. Eng.* **2014**, *2*, 1474.
3. Miao, S.; Wang, P.; Su, Z.; Zhang, S. *Acta Biomater.* **2014**, *10*, 1692.
4. Mosiewicki, M. A.; Dell'Arciprete, G. A.; Aranguren, M. I.; Marcovich, N. E. *J. Compos. Mater.* **2009**, *43*, 3057.
5. Chawla, K. K. *Mater. Sci. Eng. A.* **2012**, *557*, 2.
6. Heinen, M.; Gerbase, A. E.; Petzhold, C. L. *Polym. Degrad. Stab.* **2014**, *108*, 76.
7. Guo, C.; Zhou, L.; Lv, J. *Polym. Polym. Compos.* **2013**, *21*, 449.
8. Dasari, A.; Yu, Z. Z.; Cai, G. P.; Mai, Y. W. *Prog. Polym. Sci.* **2013**, *38*, 1357.
9. Zhang, L.; Zhang, M.; Zhou, Y.; Hu, L. *Polym. Degrad. Stab.* **2013**, *98*, 2784.
10. Zheng, X.; Wang, G.; Xu, W. *Polym. Degrad. Stab.* **2014**, *101*, 32.
11. Hu, X.; Wang, D.; Cheng, W. *Int. J. Min. Sci. Technol.* **2015**, *26*, 345.
12. Bian, X.-C.; Tang, J.-H.; Li, Z.-M. *J. Appl. Polym. Sci.* **2008**, *110*, 3871.
13. Zhang, M.; Luo, Z.; Zhang, J.; Chen, S.; Zhou, Y. *Polym. Degrad. Stab.* **2015**, *120*, 427.
14. Yang, R.; Hu, W.; Xu, L.; Song, Y.; Li, J. *Polym. Degrad. Stab.* **2015**, *122*, 102.
15. Zhao, C.; Yan, Y.; Hu, Z.; Li, L.; Fan, X. *Construct. Build Mater.* **2015**, *93*, 309.
16. Jia, D.; Hu, J.; He, J.; Yang, R. *J. Appl. Polym. Sci.* **2019**.
17. Wang, X.; Zhang, P.; Huang, Z.; Xing, W.; Song, L.; Hu, Y. *Fire Saf. J.* **2019**, *106*, 72.
18. Atabek, L.S. **2019**, *294*.
19. Shukla, D.; Srivastava, R. *Proc. World Congr. Eng.* **2011**, *III*, 3.
20. Zirconium Oxide: Properties, Uses And Manufacturing Process. https://www.articlecube.com/zirconium-oxide-properties-uses-and-manufacturing-process (accessed Sep 18, 2017)
21. Mishra, T. K.; Kumar, A.; Verma, V.; Pandey, K. N.; Kumar, V. *Compos. Sci. Technol.* **2012**, *72*, 1627.
22. Verdolotti, L.; Salerno, A.; Lamanna, R.; Nuziata, A.; Netti, P.; Lannace, S. *Micropor Mesoporous Mat.* **2012**, *151*, 79.
23. Swain, S.; Sharma, R. A.; Bhattacharya, S.; Chaudhary, L. *Trans. Electr. Electron. Mater.* **2013**, *14*, 1.
24. Xu, K.; Hu, Y. Q.; *Chinese J. Polym. Sci. (English Ed.)* **2010**, *28*, 13.
25. Zhou, S.; Wu, L. *Macromol. Chem. Phys.* **2008**, *209*, 1170.
26. Azemati, A. A.; Khorasanizadeh, H.; Shirkavand Hadavand, B.; Sheikhzadeh, G. A. *Mater. Sci. Forum* **2017**, *894*, 109.
27. Wang, Z.; Huang, Z.; Li, X.; Zhou, J. *Composite Part A Sci. Manuf.* **2019**, *121*, 180.
28. Agrawal, A.; Kaur, R.; Walia, R. S. *e-polymers* **2019**, *19*, 411.
29. Kaur, R.; Kumar, M. *J. Polym. Eng.* **2013**, *33*, 875.
30. Zieleniewska, M.; Leszczyński, M. K.; Szczepkowski, L.; Bryśkiewicz, A.; Małgorzata, K.; Bień, K.; Ryszkowska, J. *Polym. Degrad. Stab.* **2016**, *132*, 78.
31. Fan, H.; Tekeei, A.; Suppes, G. J.; Hsieh, F. H. *Int. J. Polym. Sci.* **2012**, *2012, 1*.
32. Rigid Polyurethane Foam Sheets - High Density Hard Foam Blocks | General Plastics — General Plastics. https://www.generalplastics.com/rigid-foams (accessed June 29, 2018)
33. Yang, Y.C.; Jeong, S.B.; Kim, B.G.; Yoon, P.R. *powder Technol.* **2009**, *191*, 117.

34. Wang, Y.; Wang, F.; Dong, Q.; Xie, M.; Liu, P.; Ding, Y.; Zhang, S.; Yang, M.; Zheng, G. *Polym. Degrad. Stab.* **2017**, *146*, 267.
35. Zhang, M.; Zhang, J.; Chen, S.; Zhou, Y. *Polym. Degrad. Stab.* **2014**, *110*, 27.
36. Hull, T. R.; Witkowski, A.; Hollingbery, L. *Polym. Degrad. Stab.* **2011**, *96*, 1462.
37. Yang, R.; Wang, B.; Han, X.; Ma, B.; Li, J. *Polym. Degrad. Stab.* **2017**, *144*, 62.

Fractional Order High Pass Filter Based on Operational Transresistance Amplifier with Three Fractional Capacitors of Different Order

Gagandeep KAUR[1], *Abdul Quaiyum ANSARI*[1], *Mohammad Shabi HASHMI*[2]

[1]Department of Electrical Engineering, Jamia Millia Islamia, Okhla, Jamia Nagar, Delhi 110025, India
[2]Department of Electronics and Communication Engineering, Indraprastha Institute of Information Technology (IIIT), Okhla Industrial Estate, Phase III, New Delhi 110020, India

arora_gagan07@yahoo.co.in, aqansari@jmi.ac.in, mshashmi@iiitd.ac.in

DOI: 10.15598/aeee.v17i2.2998

Abstract. *Design, realization and performance evaluation of fractional order high pass filter based on Operational Transresistance Amplifier (OTRA) using three fractional elements of different order α, β, γ are presented in this paper. The projected circuit uses a single active current mode device i.e. OTRA and utilizes fractional capacitors of varying orders to boost the design flexibility. The design equations for fractional order high pass filter are computed from the transfer function. The paper elaborates the impact of fractional order elements of varying order on the frequency response of the filter. Subsequently, the sensitivity and stability of the transfer function of the proposed filter are also analyzed. The potential usefulness of the proposed filter is additionally incontestable through the Total Harmonic Distortion (THD), Power Supply Rejection Ratio (PSRR), Temperature sweep, Corner, Supply Voltage variation and Noise Analysis. It has been observed that the proposed pass filter in fractional domain supported OTRA provides greater flexibility in controlling the magnitude characteristics.*

Keywords

Fractional High Pass Filter (FHPF), Fractional Order Element (FOE), Fractional Order Signal Processing (FOSP), Operational Transresistance Amplifier (OTRA).

1. Introduction

Signal processing applications supported by fractional order calculus are termed as Fractional Order Signal Processing (FOSP). Many applications supported by FOSP are mentioned within the literature which includes FO Oscillators [1] and [2], Filters [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17] and [18], Multivibrators [19], Differentiators and Integrators [20] and [21], etc. Active and passive integer order filters are realized by using components, such as resistor, capacitor and inductor in the time domain. Similarly, an inductor or capacitor whose impedance is given as $Z(s) = as^{\alpha} \Rightarrow Z(j\omega) = a\omega^{\alpha}e^{j(\pi\alpha/2)}$ are the kinds of fractance device which are the essential components in the realization of Fractional Order (FO) filters where $\alpha = 1, 0, -1$ represents an inductor, a resistor and a capacitor respectively in complex frequency domain [22], [23], [24] and [25]. The corresponding phase angle is $\pi/2$ and $-\pi/2$ for the impedance of the inductor and capacitor respectively. The passive elements can be therefore termed as Fractional Order Elements (FOEs) that have a constant phase angle with frequency. The impedance of this FOE is given as $Z_C^{\alpha}(s) = 1/s^{\alpha}C$ where α is the order of FO capacitance, C is the capacitance with units $F/s^{1-\alpha}$ and s is the unit of time [10].

In the design of FO systems, the realization of FO capacitors (FCs) and FO inductors (FIs) is a vital purpose of concern. Antecedently the realization of FOE approximation of type $Z = 1/Cs^{1/n}$ was reported [26], [27], [28] and [29] and these embrace realization either by RC trees [25], nested ladder [30] or domino ladder [31]. Subsequently, fabrication of two terminal fractance devices is reported in [22], [32] and [45] and these consist of metal insulator-liquid interface. Recently the fabrication of FOEs [33] and [34] using copper electrode and platinised silicon electrode is reported which has the advantage of better reproducibility and is smaller in dimensions. However, it has obstacles of a short lifetime, high cost, and is nonetheless not mass producible.

In the most recent studies, the current mode technique has been preferred in integrated circuit design because of its inherent superiority of high slew rate, wide bandwidth, low power consumption, and high speed over the voltage mode approach [35]. OTRA is a high gain current input voltage output device that acquires all the advantages of current mode techniques. OTRA is used as an active element for realizing signal processing and generation circuits resembling filters [35], [36] and [37], oscillators [38], multivibrators [39], and immitance simulators [40]. Due to varied applications of filters in instrumentation, communication and signal processing, numerous reported designs aim to generalize the design schemes of integer order filters to the fractional domain [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [41] and [42]. It can be identified from the literature that active FO filters based on FOEs are designed using Op-Amps [3], [4], [5], [7], [8], [10], [11], [16] and [42] however there are very few design reports of FO filters using current mode active elements [9], [12], [13], [14], [15] and [41]. There are reports of some integer order filters based on OTRA using one or more active components [35], [36] and [37], but high pass filter using single OTRA in the fractional domain has not nevertheless been reported.

In this paper, therefore, design procedure, realization and performance evaluation fractional of third order high pass filter using single OTRA is reported. The proposed circuit has the following characteristics:

- The proposed circuit uses single active current mode device i.e. OTRA, and utilizes fractional capacitors of different orders to enhance the design flexibility.
- The proposed circuit has a low RMS value of noise voltage and THD at the output.
- The proposed circuit has high PSRR.

The next section presents the proposed design of the fractional high pass filter for various values of order α, β, and γ. The stability and sensitivity analysis of the presented filter is additionally delineated in Sec. 2. The evaluation of the proposed design is mentioned in Sec. 3. while Sec. 4. concludes the paper.

2. Proposed Design Procedure of High Pass Filter of Order $(\alpha + \beta + \gamma)$ Using OTRA

2.1. Circuit Description

The voltage and current relationship of the OTRA, shown symbolically in Fig. 1, is characterised by Eq. (1) where R_m is the transresistance gain of OTRA. The Current Feedback Amplifier (CFA) primarily based OTRA given in Fig. 2 is employed to realize high pass fractional filter of order $(\alpha + \beta + \gamma)$ shown in Fig. 3 [37]. It is apparent that it uses single OTRA, four resistances, and five FOC's.

$$\begin{bmatrix} V_p \\ V_n \\ V_0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ R_m & -R_m & 0 \end{bmatrix} \begin{pmatrix} I_p \\ I_n \\ I_0 \end{pmatrix}. \tag{1}$$

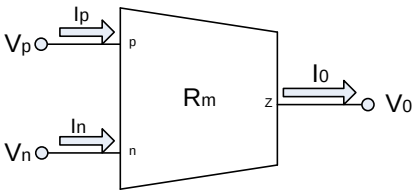


Fig. 1: Symbol of OTRA.

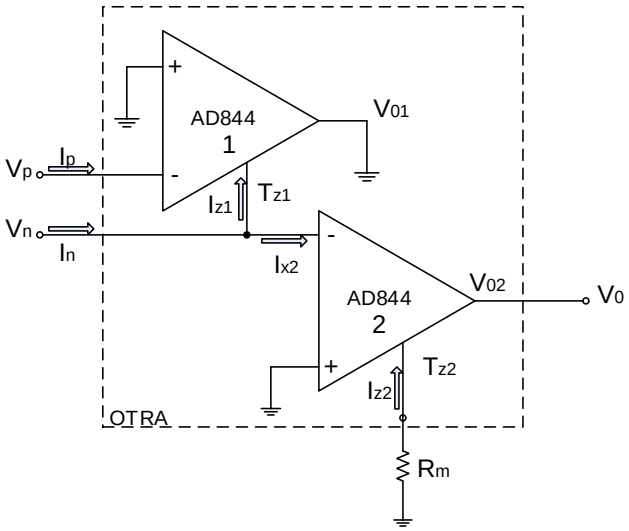


Fig. 2: CFA based OTRA.

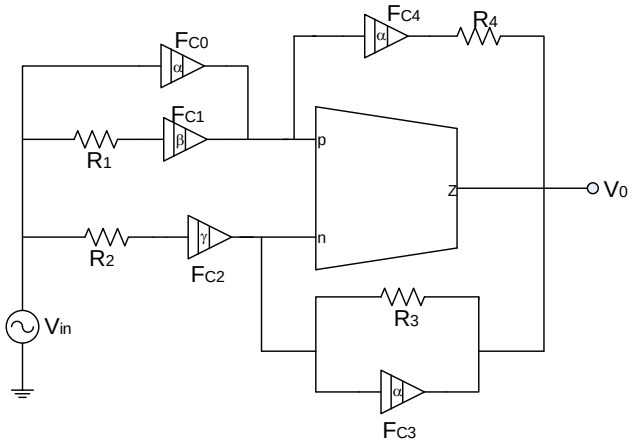


Fig. 3: Fractional order high pass filter based on OTRA.

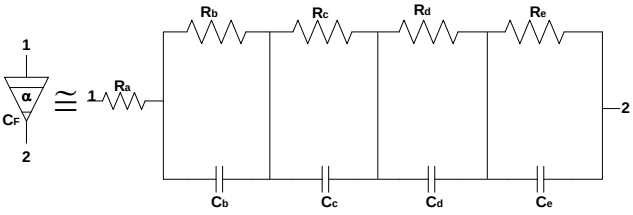


Fig. 4: Fourth order approximated representation of fractional capacitor.

The transfer function, given in Eq. (2), of the filter in Fig. 3 can be derived using KCL and nodal method. The simplification of the transfer function has been carried out under the following assumption:

- $F_{C0} = F_{C3}$,
- $F_{C1}F_{C2}R_2R_3 + F_{C0}F_{C2}R_2R_3 + F_{C0}F_{C1}R_1R_3 = F_{C1}F_{C2}R_1R_3$,
- $F_{C2}R_2 = F_{C4}R_4$,
- $F_{C0} + F_{C1} = F_{C2}$.

$$H(s) = \frac{ks^{\alpha+\beta+\gamma}}{s^{\alpha+\beta+\gamma} + d_2 s^{\alpha+\beta} + d_1 s^{\alpha} + d_0}, \tag{2}$$

where $d_2 = \left(\frac{1}{F_{C2}R_3} + \frac{1}{F_{C2}R_2} + \frac{1}{F_{C1}R_1} - \frac{1}{F_{C3}R_2}\right)$, $d_1 = \left(\frac{1}{F_{C2}F_{C3}R_2R_3} + \frac{1}{F_{C1}F_{C2}R_1R_2} + \frac{1}{F_{C1}F_{C2}R_3R_1} + -\frac{1}{F_{C1}F_{C3}R_1R_2}\right)$, $d_0 = \frac{1}{F_{C1}F_{C2}F_{C3}R_1R_2R_3}$ and $k = 1$.

The magnitude and phase responses of the presented filter are given in Eq. (3) and Eq. (4):

$$|H(j\omega)| = \frac{k\omega^{\alpha+\beta+\gamma}}{\left[\begin{array}{c} \omega^{\alpha+\beta+\gamma} + 2d_2\cos\left(\frac{\gamma\pi}{2}\right)\omega^{2(\alpha+\beta)+\gamma} \\ +2d_1\cos\left(\frac{(\beta+\gamma)\pi}{2}\right)\omega^{2\alpha+\beta+\gamma} + \\ d_2^2\omega^{2(\alpha+\beta)} + 2d_2d_1\cos\left(\frac{\beta\pi}{2}\right)\omega^{2\alpha+\beta} \\ +2d_0\cos\left(\frac{(\alpha+\beta+\gamma)\pi}{2}\right)\omega^{\alpha+\beta+\gamma} \\ +2d_2d_0\cos\left(\frac{(\alpha+\beta)\pi}{2}\right)\omega^{\alpha+\beta} \\ +d_1^2\omega^{2\alpha} + 2d_1d_0\cos\left(\frac{\alpha\pi}{2}\right)\omega^{\alpha} + d_0^2 \end{array}\right]^{\frac{1}{2}}}, \tag{3}$$

$$\arg\{H(j\omega)\} = \arg\{k\} + \frac{(\alpha+\beta+\gamma)\pi}{2} + - \tan^{-1}\left(\frac{\begin{array}{c} \omega^{\alpha+\beta+\gamma}\sin\left(\frac{(\alpha+\beta+\gamma)\pi}{2}\right) + \\ d_2\omega^{\alpha+\beta}\sin\left(\frac{(\alpha+\beta)\pi}{2}\right) + \\ d_1\omega^{\alpha}\sin\left(\frac{\alpha\pi}{2}\right) \end{array}}{\begin{array}{c} \omega^{\alpha+\beta+\gamma}\cos\left(\frac{(\alpha+\beta+\gamma)\pi}{2}\right) + \\ d_2\omega^{\alpha+\beta}\cos\left(\frac{(\alpha+\beta)\pi}{2}\right) + \\ d_1\omega^{\alpha}\cos\left(\frac{\alpha\pi}{2}\right) + d_0 \end{array}}\right). \tag{4}$$

The phase of FHP filter is equal to $\left(\arg\{k\} + \frac{(\alpha+\beta+\gamma)\pi}{2}\right)$ for $\omega \to 0$ where the phase is equal to $(\arg\{k\})$ for $\omega \to \infty$. The critical frequencies, the peak value frequency ω_p (the maximum and minimum of magnitude occurs at this frequency), cut-off frequency ω_c (the power falls to $1/\sqrt{2}$ of passband at this frequency) [3], [4], [5] and [9] and right phase frequency ω_r are the important parameters of a filter. The peak value frequency ω_p of the high pass filter can be computed by solving the condition $\frac{\mathrm{d}}{\mathrm{d}\omega}|H(j\omega)|_{\omega=\omega_p} = 0$ and is given in Eq. (5):

$$\begin{array}{c} d_2\gamma\cos\frac{\gamma\pi}{2}\omega_p^{2\alpha+2\beta+\gamma} + d_1(\beta+\gamma)\cos\frac{(\beta+\gamma)\pi}{2}\omega_p^{2\alpha+\beta+\gamma} \\ +d_2^2\gamma\omega_p^{2\alpha+2\beta} + d_1d_2(\beta+2\gamma)\cos\frac{\beta\pi}{2}\omega_p^{2\alpha+\beta} \\ +d_0(\alpha+\beta+\gamma)\cos\frac{(\alpha+\beta+\gamma)\pi}{2}\omega_p^{\alpha+\beta+\gamma} \\ +d_1^2(\beta+\gamma)\omega_p^{2\alpha} \\ +d_0d_2(\alpha+\beta+2\gamma)\cos\frac{((\alpha+\beta)\pi}{2}\omega_p^{\alpha+\beta} \\ +d_0d_1(\alpha+2\beta+2\gamma)\cos\frac{\alpha\pi}{2}\omega_p^{\alpha} \\ +d_0^2(\alpha+\beta+\gamma) = 0. \end{array} \tag{5}$$

The cut-off frequency ω_c of the presented high pass filter can be calculated through the condition $|H(j\omega)|\omega = \omega_c = 0.707 \cdot |H(j\omega)|_{\max}$ and is given in Eq. (6):

$$\begin{array}{c} \omega_c^{2(\alpha+\beta+\gamma)} - 2d_1d_2\cos\frac{\gamma\pi}{2}\omega_c^{2(\alpha+\beta)+\gamma} - \\ 2d_1\cos\frac{(\gamma+\beta)\pi}{2}\omega_c^{2\alpha+(\beta+\gamma)} - d_2^2\omega_c^{2(\alpha+\beta)} - \\ 2d_1d_2\cos\frac{\beta\pi}{2}\omega_c^{2\alpha+\beta} - 2d_0\cos\frac{(\alpha+\beta+\gamma)\pi}{2}\omega_c^{\alpha+\beta+\gamma} \\ -2d_0d_2\cos\frac{(\alpha+\beta)\pi}{2}\omega_c^{\alpha+\beta} - d_1^2\omega_c^{2\alpha} \\ -2d_0d_1\cos\frac{\alpha\pi}{2}\omega^{\alpha} - d_0^2 = 0. \end{array} \tag{6}$$

The right phase frequency ω_r at which the phase $\angle H(j\omega_r) = \pm\frac{\pi}{2}$ for high pass filter can be acquired from the following Eq. (7):

$$\omega_r^{\alpha+\beta+\gamma} + d_2\omega_r^{\alpha+\beta}\cos\frac{\gamma\pi}{2} + d_1\omega_r^{\alpha}\cos\frac{(\beta+\gamma)\pi}{2} + \\ +d_0\cos\frac{(\alpha+\beta+\gamma)\pi}{2} = 0. \tag{7}$$

The theoretical results obtained from Eq. (3) and Eq. (4) at $d_2 = 2$, $d_1 = 2$, and $d_0 = 1$ for different values of α, β, γ are plotted in Fig. 5. The value of d_2, d_1, and d_0 are taken from the Butterworth filter coefficient table for the third order filter. It is observed from the simulations that the magnitude and phase response of the proposed filter are not only controlled with the coefficients (d_2, d_1, d_0) but can also be controlled with fractional order (α, β, γ) which increases the design flexibility of the proposed filter.

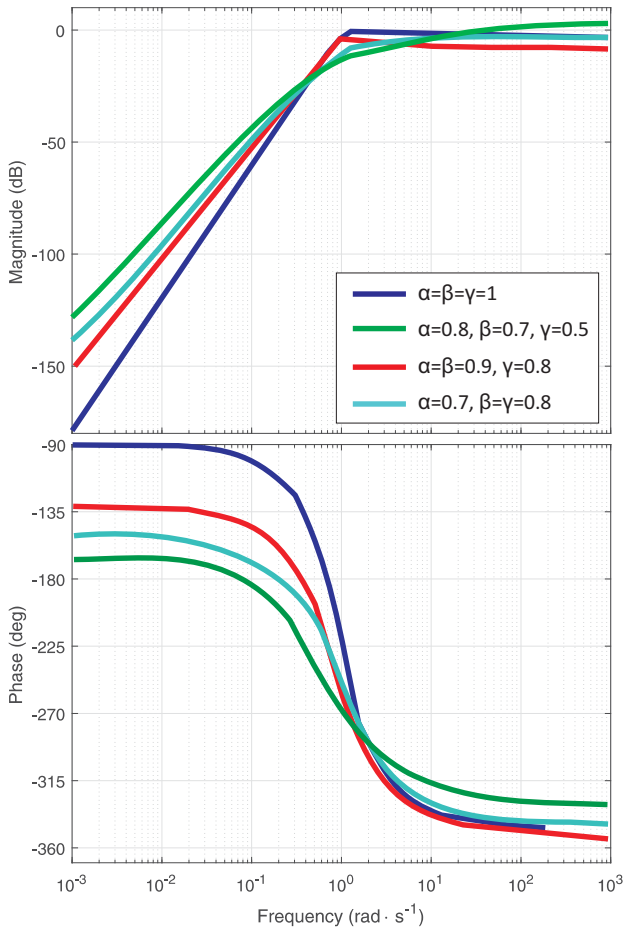


Fig. 5: MATLAB simulations of magnitude and phase response of proposed filter for different values of α, β and γ.

2.2. Non-Ideal Analysis

The performance of the FHPF may deviate because of non-idealities of OTRA in practice. Ideally OTRA is an infinite gain ($R_m \to \infty$), differential Current Controlled Voltage Source (CCVS) device. However, practically the value of R_m is a frequency dependent.

Considering a single pole model of an OTRA, R_m can be expressed as:

$$R_m(s) = \left[\frac{R_d}{1+\frac{s}{\omega_0}}\right], \tag{8}$$

where R_d represents the dc transresistance gain. For high frequency applications, $R_m(s)$ reduces to:

$$R_m(s) \approx \frac{1}{sC_p}, \text{ where } C_p = \frac{1}{R_d\omega_0}.$$

Taking this consequence into account the transfer function given in Eq. (2) transforms to Eq. (9).

Where the uncompensated error term sC_p appearing in parallel to F_{C3} can lead to the introduction of another pole. The value of F_{C3} may be decreased by C_p to absorb the nonideality effect and thus achieving self-compensation. Another nonideality present in it is because of parasitic resistances and capacitances at higher frequencies shown in Fig. 6(a).

2.3. Stability

It is an important property of a fractional filter that can be investigated by converting the fractional domain into various domains [43]. The stability of the proposed fractional filter is studied by converting the s-domain into a W-domain that transforms the fractional order transfer function into integer order. The following steps are performed for the stability analysis of this filter:

- Convert the fractional domain into W-domain by using $W = s^{1/m}$ and $\alpha = k_1/m$ and $\beta = k_2/m$ and by selecting the values of k_1, k_2, and m for the desired value of α and β.
- Solve the converted transfer function to compute minimum root angle$|\theta_w|$. If the minimum root angle$|\theta_w|$ is less than $\pi/2m$ then the filter is unstable otherwise it is stable.

This procedure is then applied to the denominator of Eq. (2) that gives the characteristic equation in the W-plan.

$$H(s)=\frac{s^{\alpha+\beta+\gamma}}{s^{\alpha+\beta+\gamma}+\left(\frac{1}{F_{C2}R_3}+\frac{1}{F_{C2}R_2}+\frac{1}{F_{C1}R_1}-\frac{1}{(F_{C3}+sC_p)R_2}\right)s^{\alpha+\beta}+\ldots \atop \ldots+\left(\frac{1}{F_{C2}(F_{C3}+sC_p)R_2R_3}+\frac{1}{F_{C1}F_{C2}R_1R_2}+\frac{1}{F_{C1}F_{C2}R_3R_1}-\frac{1}{F_{C1}(F_{C3}+sC_p)R_1R_2}\right)s^{\alpha}\ldots \atop \ldots+\frac{1}{F_{C1}F_{C2}(F_{C3}+sC_p)R_1R_2R_3}}. \tag{9}$$

$$W^{k_1+k_2+k_3}+d_2W^{k_1+k_2}+d_1W^{k_1}+d_0. \tag{10}$$

The roots of the Eq. (10) were calculated for various values of α, β, and γ with $d_2=2$, $d_1=2$, and $d_0=1$. It was observed that the minimum root angle $|\theta_w|_{\min}>\pi/2q=9^\circ$ where $q=10$ for various values of α, β and γ. The stability of the filter for various values of α and β with coefficients d_2, d_1, and d_0 is given in Tab. 1 which confirms that the filter using these coefficients is stable when $\alpha\leq 1$, $\beta\leq 1$ and $\gamma\leq 1$.

Tab. 1: Minimum root angle for the third order filter in fractional domain for different values of α, β, and γ with coefficients, $d_2=2$, $d_1=2$, and $d_0=1$.

α	β	γ	$\lvert\theta_w\rvert_{\min}$	**Stability**
0.9	0.6	0.5	15.45°	Stable
0.7	0.8	0.8	14.94°	Stable
0.9	0.9	0.8	13.29°	Stable
0.5	0.6	0.9	14.11°	Stable
0.8	0.7	0.6	16.26°	Stable
1	1	1	12.00°	Stable
1.2	0.8	0.7	11.61°	Stable
1.4	1.4	0.8	8.53°	Unstable
1.2	1.6	1.3	7.76°	Unstable

2.4. Sensitivity Analysis

Sensitivity analysis is one of the most important parameters in computing the performance of active filters and is defined as:

$$S_x^y=\left(\frac{x}{y}\right)\left(\frac{\delta x}{\delta y}\right). \tag{11}$$

It means that the fractional change in the parameter (y) i.e. transfer function is normalised by the fractional change in component (x) value. Change in the sensitivity of transfer function relative to component variation is complex and difficult to compute [9]. The transfer function sensitivity relative to fractional orders, i.e. sensitivity toward α, β, γ can be derived by applying Eq. (11) to the transfer function of the proposed high pass filter, the sensitivities are given as:

$$S_\alpha^{T_{FHP}}=\frac{\alpha\ln s}{(s^{\alpha+\beta+\gamma}+d_2s^{\alpha+\beta}+d_1s^{\alpha}+d_0)}. \tag{12}$$

$$S_\beta^{T_{FHP}}=\frac{\beta(d_1s^{\alpha}+d_0)ln(s)}{(s^{\alpha+\beta+\gamma}+d_2s^{\alpha+\beta}+d_1s^{\alpha}+d_0)}. \tag{13}$$

$$S_\gamma^{T_{FHP}}=\frac{\gamma(d_2s^{\alpha+\beta}+d_1s^{\alpha}+d_0)ln(s)}{(s^{\alpha+\beta+\gamma}+d_2s^{\alpha+\beta}+d_1s^{\alpha}+d_0)}. \tag{14}$$

For the special case of $\alpha=\beta=\gamma=1$, the natural frequency ω_0 and quality factor Q_0 of the proposed filter can be obtained by using normal capacitors is given as:

$$\omega_0=\frac{1}{\sqrt[3]{C_1C_2C_3R_1R_2R_3}}. \tag{15}$$

$$Q_0=\frac{\sqrt[3]{C_1C_2C_3R_1R_2R_3}}{C_1R_1+C_2R_2+C_3R_3-C_2R_3-\sqrt[3]{C_1C_2C_3R_1R_2R_3}}. \tag{16}$$

The sensitivities of ω_0 w.r.t. various passive components are given as:

$$S_{C_1}^{\omega_0}=S_{C_2}^{\omega_0}=S_{C_3}^{\omega_0}=S_{R_1}^{\omega_0}=S_{R_2}^{\omega_0}=S_{R_3}^{\omega_0}=-\frac{1}{3}. \tag{17}$$

The sensitivities of Q_0 w.r.t. various passive components are given as:

$$\begin{aligned}S_{C_1}^{Q_0}=-\frac{1}{3},S_{C_2}^{Q_0}=\frac{2}{3},S_{C_3}^{Q_0}=-\frac{1}{3},\\ S_{R_1}^{Q_0}=-\frac{1}{3},S_{R_2}^{Q_0}=-\frac{1}{3},S_{R_3}^{Q_0}=-\frac{2}{3}.\end{aligned} \tag{18}$$

It shows that the proposed circuit offers low passive sensitivity.

3. Circuit Simulation

The practicality of the proposed circuit in Fig. 3 is verified on the PSpice platform for four cases i.e.

- $\alpha\neq\beta\neq\gamma$ using factional capacitors of different order where $\alpha=0.8$, $\beta=0.7$ and $\gamma=0.5$.
- $\alpha\neq\beta=\gamma$ using two fractional capacitors of same order where $\alpha=0.7$ and $\beta=\gamma=0.8$.

Tab. 2: Component values used to realize fractional capacitor at a centre frequency of 1000 Hz.

Passive Components	Fractional Capacitor											
	$F_{c_0}=F_{c_3}=10\cdot10^{-9}[F/s^{1-\alpha}]$			$F_{c_1}=10\cdot10^{-9}[F/s^{1-\beta}]$			$F_{c_2}=20\cdot10^{-9}[F/s^{1-\gamma}]$			$F_{c_4}=5\cdot10^{-9}[F/s^{1-\alpha}]$		
	α=0.8	α=0.7	α=0.9	β=0.7	β=0.8	β=0.9	γ=0.5	γ=0.8	γ=0.7	α=0.8	α=0.7	α=0.9
R_a (Ω)	1681.73	8135.32	259.58	8135.32	1681.73	259.58	70082.47	840	4067.66	3363.47	16270.64	519.16
R_b (kΩ)	8.601	31.93	1.77	31.93	8.601	1.77	158.73	4.30	15.973	17.20	63.85	3.55
R_c (kΩ)	19.55	66.31	3.57	66.31	19.55	3.57	238.85	9.77	33.155	39.10	132.62	7.14
R_d (kΩ)	95.02	241.83	29.06	241.83	95.02	29.06	560.65	47.51	120.91	190.04	483.66	58.11
R_e (kΩ)	4853.26	5568.70	5574.59	5568.70	4853.26	5574.59	4648.35	2426.63	2784.35	9706.53	11137.4	11491.83
C_b (nF)	2.98	0.79	16.9	0.79	2.98	16.9	0.132	5.97	1.59	1.49	0.4	2.72
C_c (nF)	3.21	1.66	23.57	1.66	3.21	23.57	0.469	6.42	3.29	1.60	0.83	8.45
C_d (nF)	6.54	1.96	42.44	1.96	6.54	42.44	0.851	13.08	4.70	3.27	0.98	11.78
C_e (nF)	7.18	2.35	54.4	2.35	7.18	54.4	1.10	14.36	3.31	3.59	1.17	21.22

Tab. 3: Relative study of proposed filter with other fractional filters available in the literature.

Ref.	Type of Fractional filter	Circuit simulations/ Experimental work of fractional filter	Type and number of active components used	Number of components used			Performance Parameter					
				R	C	FOC	Critical Frequencies	Sensitivity Analysis	Stability Analysis	Noise Analysis	THD	PSRR
[7]	KHN	FLPF of order $(\alpha+\beta)$, at $\alpha=0.7$, $\beta=1.2$ and $\alpha=0.7$, $\beta=0.7$	Op-amp,3	6	-	2	Yes	No	Yes	No	No	No
	Sallen-key	FLPF of order $(\alpha+\beta)$, at $\alpha=1.2$, $\beta=1.5$ and $\alpha=1.2$, $\beta=0.7$	Op-amp,1	4	-	2						
[8]	KHN	FLPF, FHPF and FBPF of order $(\alpha+\beta)$ at $\alpha=0.86$, $\beta=0.46$	Op-amp,3	6	-	2	Yes	Yes	Yes	No	No	No
[9]	KHN	FLPF of order $(\alpha+\beta)$ at $\alpha=1.2$, $\beta=0.4$; $\alpha=0.8$, $\beta=0.7$ and $\alpha=1.2$, $\beta=1.7$	CCII,4	6	-	2	Yes	No	Yes	No	No	No
	Tow-Tomas	FLPF of order $(\alpha+\beta)$ at $\alpha=1.2$, $\beta=0.4$; $\alpha=1.7$, $\beta=1.2$ and $\alpha=0.8$, $\beta=0.7$	CCII,3	4	-	2						
[10]	FO Chebyshev LPF	FLPF of order $(1+\alpha)$ at $\alpha=0.2$, 0.5 and 0.8	Op-amp,3	6	1	1	No	No	Yes	No	No	No
[11]	FO inverse Chebyshev LPF	FLPF of order $(1+\alpha)$ at $\alpha=0.2$, 0.5 and 0.8	Op-amp,3	6	1	1	No	No	Yes	No	No	No
[12]	FHPF with electronically controllable parameter	FHPF of order $(1+\alpha)$ at $\alpha=0.25$, 0.5 and 0.75	OTA,3 ACA,2 DO-CF,1	-	3	-	Yes	No	No	No	No	No
[13]	FLPF	FLPF of order 2α at $\alpha=0.8$	CCII,1	2	-	2	Yes	Yes	Yes	No	No	No
[14]	FLPF with electronic tunability	FLPF of order $(1+\alpha)$ at $\alpha=0.1$, 0.3, 0.5, 0.7 and 0.9	OTA,2 ACA,2 and MO-CF,1	-	3	-	Yes	No	No	No	Yes	No
[15]	CDBA based fractional order Multi function filter	FLPF, FHPF, FBPF at $\alpha=\beta=\gamma=1$ $\alpha=0.5$,$\beta=0.9$,$\gamma=0.7$ $\alpha=\beta=\gamma=0.9$	CDBA,1	5	-	5	Yes	Yes	Yes	Yes	Yes	No
Proposed	FHPF	FHPF of order $(\alpha+\beta+\gamma)$ at $\alpha=\beta=\gamma=1$ $\alpha=\beta=0.9$, $\gamma=0.8$ $\alpha=0.7$, $\beta=\gamma=0.8$ $\alpha=0.8$, $\beta=0.7$, $\gamma=0.5$	OTRA,1	4	-	5	Yes	Yes	Yes	Yes	Yes	Yes

- $\alpha=\beta\neq\gamma$ using two fractional capacitors of same order where $\alpha=\beta=0.9$ and $\gamma=0.8$.
- $\alpha=\beta=\gamma=1$ using a normal capacitor.

In all the cases the circuit remains in the stable region given in Tab. 3. The component values to obtain the frequency response of the proposed filter are:

- $F_{c_0}=10\cdot10^{-9}[F/s^{\alpha}]$,
- $F_{c_1}=10\cdot10^{-9}[F/s^{\beta}]$,
- $F_{c_2}=20\cdot10^{-9}[F/s^{\gamma}]$,
- $F_{c_3}=10\cdot10^{-9}[F/s^{\alpha}]$,
- $F_{c_4}=5\cdot10^{-9}[F/s^{\alpha}]$,
- $R_1=10$ kΩ, $R_2=2.5$ kΩ, $R_3=20$ kΩ, $R_4=10$ kΩ.

The fractional capacitors used for the Pspice simulations are realized as shown in Fig. 4 by fourth order rational approximation for s^{α} [44] using Continued Fraction Expansions (CFEs). By using this

rational approximation, the fractional Laplacian operator can be realised physically with the RC ladder network. The impedance (Z) of this RC network is:

$$Z = R_a + \cfrac{\frac{1}{C_b}}{s + \cfrac{1}{R_b C_b}} + \cfrac{\frac{1}{C_c}}{s + \cfrac{1}{R_c C_c}} + \cfrac{\frac{1}{C_d}}{s + \cfrac{1}{R_d C_d}} + \cfrac{\frac{1}{C_e}}{s + \cfrac{1}{R_e C_e}}. \tag{19}$$

The values of resistances and capacitors are obtained by equating terms of fourth order rational approximation with the terms of the Eq. (19) and are given in Tab. 2 designed at a centre frequency of 1000 Hz. The PSpice simulations of Fig. 3 are shown in Fig. 6(a). It is observed that the simulation results are in accordance with theoretical results. Table 3 shows the relative study of the proposed filter with the fractional filters available in the literature. Noise analysis is also performed on the proposed fractional order high pass filter based on OTRA with three degrees of freedom using PSpice AC Sweep/Noise simulation program. Figure 6(b) shows the noise voltage and the Root Mean Square (RMS) value of the noise voltage at the output. The RMS value of noise voltage at the output of the proposed filter for all the four cases are $1.4 \cdot 10^{-7}$ V for order 2, $1.62 \cdot 10^{-7}$ V for order 2.3, $1.97 \cdot 10^{-7}$ V for order 2.6 and $3.89 \cdot 10^{-7}$ V for order 3. It is observed that the RMS value of noise voltage at the output of the fractional filter is less than that of the integer order filter. THD values are derived for input voltage amplitude values from 0 mV to 4000 mV from PSPICE Transient and Fourier analysis and are given in Fig. 6(c). The THD analysis results show that when the input voltage is below 2000 mV, the THD is less than 1 %: THD is less than 5 % of input voltage is below 4000 mV for all the four cases.

For PSRR measurement of the projected fractional order high pass filter, DC bias is connected to the input and AC signal is connected at VDD terminal. The PSRR analysis shows that the proposed circuit provides a 69.643 dB for order 2, +66.640 dB for order 2.3, 64.681 dB for order 2.6 and 59.760 dB for order 3 at 10 kHz. The simulation results for PSRR given in Fig. 6(d) show that the fractional order circuits have higher PSRR when compared to the integer order ones under similar conditions. Temperature sweep analysis is also performed on the proposed fractional order high pass filter based on OTRA with three degrees of freedom using PSPICE Temperature Sweep/transient analysis. Figure 6(e) shows the output voltage variation with change in temperature ranging from −50 °C to +50 °C. It is observed from the simulations that as the temperature increases the output of the circuit decreases. Figure 6(f) shows the change in output voltage with respect to supply voltage variation from 0 to 10 V at different values of α, β and γ. The simulation shows that the output voltage varies linearly with the supply voltage.

Corner analysis of the proposed circuit is given in Fig. 7(a) at $\alpha = 0.8$, $\beta = 0.7$, $\gamma = 0.5$, Fig. 7(b) at $\alpha = 0.7$, $\beta = \gamma = 0.8$ Fig. 7(c) at $\alpha = \beta = 0.9$, $\gamma = 0.8$ and Fig. 7(d) at $\alpha = \beta = \gamma = 1$. Three corners typical-typical (tt), slow-slow (ss) and fast-fast (ff) are considered. DC gain indicates approximately 3 % (ff) and 7.1 % (ss) change from the (tt) value for all the cases, therefore (ss) corner can be considered a worst case for this circuit. The performance parameters of the planned filter at different values of α, β and γ are tabulated in Tab. 4.

Tab. 4: Performance parameters of proposed fractional order High Pass Filter.

Order of fractional order HPF	Values of α β and γ	RMS value of Noise Voltage (V)	THD (%) @4 mV	PSRR (dB) @ 10 kHz
2.0	$\alpha = 0.8$ $\beta = 0.7$ $\gamma = 0.5$	$1.4 \cdot 10^{-7}$	3.51	69.643
2.3	$\alpha = 0.7$ $\beta = 0.8$ $\gamma = 0.8$	$1.62 \cdot 10^{-7}$	3.60	66.640
2.6	$\alpha = 0.9$ $\beta = 0.9$ $\gamma = 0.8$	$1.97 \cdot 10^{-7}$	3.95	64.681
3.0	$\alpha = 1$ $\beta = 1$ $\gamma = 1$	$3.89 \cdot 10^{-7}$	4.50	59.760

4. Conclusion

In this work, a fractional high pass filter based on OTRA with three fractional capacitors of different order has been realized. Mathematical equations have been considered to develop the design procedure for high pass fractional filter. The stability and sensitivity analysis in the fractional domain has been conferred as well. The Noise Analysis shows that the RMS value of noise voltage at the output of the proposed filter for all the four cases is $1.4 \cdot 10^{-7}$ V for order 2, $1.62 \cdot 10^{-7}$ V for order 2.3, $1.97 \cdot 10^{-7}$ V for order 2.6 and $3.89 \cdot 10^{-7}$ V for order 3 and these are relatively low in comparison to the other reported designs. It is conjointly identified that the RMS value of noise voltage and THD at the output is low, whereas PSRR is at higher level for fractional order circuits as compared to the corresponding integer order circuit. Supply voltage variation analysis shows that the output voltage varies linearly with supply voltage. The simulated results validate the working principle of the OTRA based high pass filter in the fractional domain.

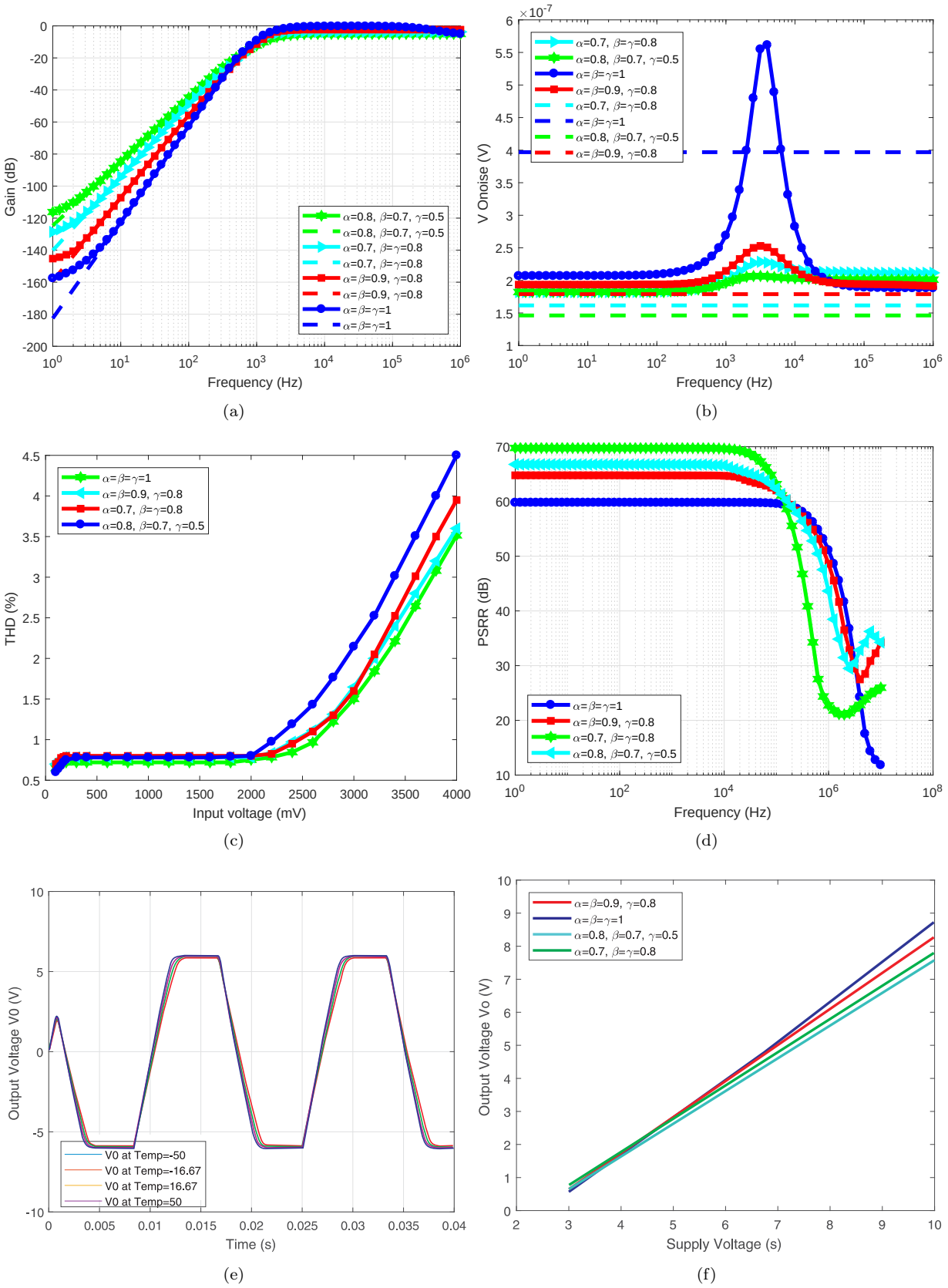
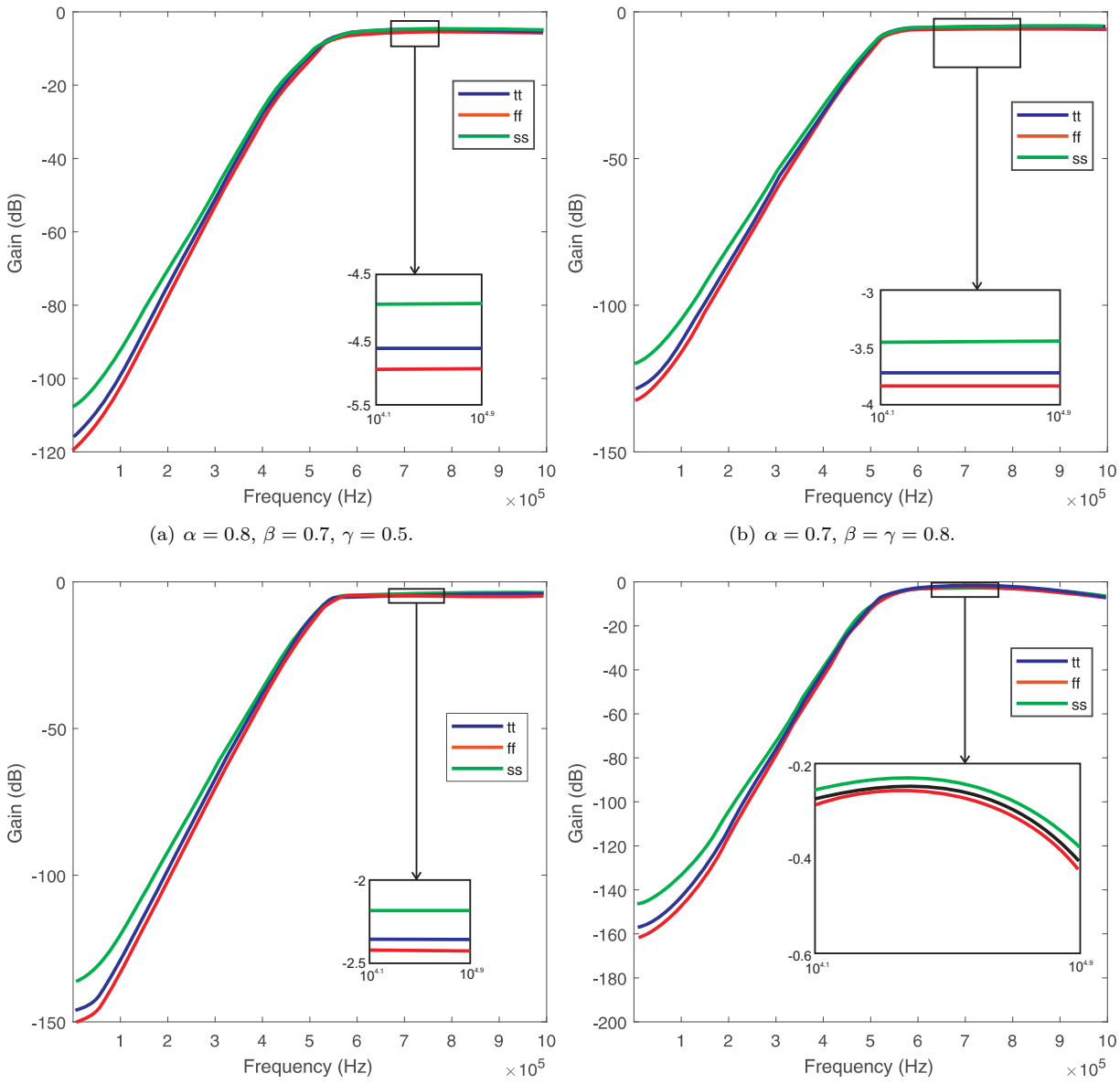


Fig. 6: (a) PSPICE simulation results of FHPF at different values of α, β and γ (dotted lines represent the theoretical response and simulated response is represented by solid lines). (b) Output noise analysis of FHPF at different values of α, β and γ (dotted lines represent the RMS value of the output noise voltage and the maximum value of output noise voltage is represented by solid lines) and (c) change of percentage THD with respect to the input voltage of FHPF at different values of α, β and γ. (d) PSRR of FHPF at different values of α, β and γ. (e) The temperature sweep of FHPF at $\alpha = \beta = 0.9$ and $\gamma = 0.8$. (f) Supply voltage variation analysis of FHPF at different values of α, β and γ.



(a) $\alpha = 0.8$, $\beta = 0.7$, $\gamma = 0.5$.

(b) $\alpha = 0.7$, $\beta = \gamma = 0.8$.

(c) $\alpha = \beta = 0.9$, $\gamma = 0.8$.

(d) $\alpha = \beta = 0.7\gamma = 1$.

Fig. 7: Corner analysis results of the FHPF.

References

[1] RADWAN, A. G., A. S. ELWAKIL and A. M. SOLIMAN. Fractional-order sinusoidal oscillators: design procedure and practical examples. *IEEE Transactions on Circuits and Systems I: Regular Papers*. 2008, vol. 55, iss. 7, pp. 2051–2063. ISSN 1549-8328. DOI: 10.1109/TCSI.2008.918196.

[2] SAID, L. A., A. G. RADWAN, A. H. MADIAN and A. M. SOLIMAN. Fractional order oscillators based on operational transresistance amplifiers. *AEU-International Journal of Electronics and Communications*. 2015, vol. 69, no. 7, pp. 988–1003. ISSN 1434-8411. DOI: 10.1016/j.aeue.2015.03.003.

[3] SOLTAN, A., A. G. RADWAN and A. M. SOLIMAN. Fractional order filter with two fractional elements of dependant orders. *Microelectronics Journal*. 2012, vol. 43, no. 11, pp. 818–827. ISSN 1879-2391. ODI: 10.1016/j.mejo.2012.06.009.

[4] RADWAN, A. G., A. M. SOLIMAN and A. S. ELWAKIL. First-order filters generalized to the fractional domain. *Journal of Circuits, Systems, and Computers*. 2008, vol. 17, no. 1, pp. 55–66. ISSN 1793-6454. DOI: 10.1142/S0218126608004162.

[5] RADWAN, A. G., A. S. ELWAKIL and A. M. SOLIMAN. On the generalization of second-order filters to the fractional-order domain. *Journal of Circuits, Systems, and Computers*. 2009, vol. 18, no. 2, pp. 361–386. ISSN 1793-6454. DOI: 10.1142/S0218126609005125.

[6] KHATEB, F., D. KUBANEK, G. TSIRIMOKOU and C. PSYCHALINOS. Fractional-order filters based on low-voltage DDCCs. *Microelectronics Journal*. 2016, vol. 50, iss. 1, pp. 50–59. ISSN 1879-2391. DOI: 10.1016/j.mejo.2016.02.002.

[7] ALI, A. S., A. G. RADWAN and A. M. SOLIMAN. Fractional order Butterworth filter: active and passive realizations. *IEEE Journal on emerging and selected topics in circuits and systems*. 2013, vol. 3, no. 3, pp. 346–354. ISSN 2156-3365. DOI: 10.1109/JETCAS.2013.2266753.

[8] TRIPATHY, M. C., K. BISWAS and S. SEN. A design example of a fractional-order Kerwin-Huelsman-Newcomb biquad filter with two fractional capacitors of different order. *Circuits, Systems, and Signal Processing*. 2013, vol. 32, no. 4, pp. 1523–1536. ISSN 1531-5878. DOI: 10.1007/s00034-012-9539-2.

[9] SOLTAN, A., A. G. RADWAN and A. M. SOLIMAN. CCII based fractional filters of different orders. *Journal of advanced research*. 2014, vol. 5, no. 2, pp. 157–164. ISSN 2090-1232. DOI: 10.1016/j.jare.2013.01.007.

[10] FREEBORN, T., B. MAUNDY and A. S. ELWAKIL. Approximated Fractional Order Chebyshev Lowpass Filters. *Mathematical Problems in Engineering*. 2015, vol. 2015, no. 832468, pp. 1–7. ISSN 1024-123X. DOI: 10.1155/2015/832468.

[11] FREEBORN, T. J., A. S. ELWAKIL and B. MAUNDY. Approximated fractional-order inverse Chebyshev lowpass filters. *Circuits, Systems, and Signal Processing*. 2016, vol. 35, no. 6, pp. 1973–1982. ISSN 0278-081X. DOI: 10.1007/s00034-015-0222-2.

[12] JERABEK, J., R. SOTNER, J. DVORAK, L. LANGHAMMER and J. KOTON. Fractional-order high-pass filter with electronically adjustable parameters. In: *2016 International Conference on Applied Electronics (AE)*. Pilsen: IEEE, 2016, pp. 111–116. ISBN 978-80-261-0602-9. DOI: 10.1109/AE.2016.7577253.

[13] SAID, L. A., S. M. ISMAIL, A. G. RADWAN, A. H. MADIAN, M. F. ABU ELYAZEED and A. M. SOLIMAN. On the optimization of fractional order low-pass filters. *Circuits, Systems, and Signal Processing*. 2016, vol. 35, no. 6, pp. 2017–2039. ISSN 0278-081X. DOI: 10.1007/s00034-016-0258-y.

[14] LANGHAMMER, L., J. DVORAK, J. JERABEK, J. KOTON and R. SOTNER. Fractional-order low-pass filter with electronic tunability of its order and pole frequency. *Journal of Electrical Engineering*. 2018, vol. 69, no. 1, pp. 3–13. ISSN 1339-309X. DOI: 10.1515/jee-2018-0001.

[15] KAUR, G., A. Q. ANSARI and M. S. HASHMI. Fractional order multifunction filter with 3 degrees of freedom. *AEU-International Journal of Electronics and Communications*. 2017, vol. 82, iss. 1, pp. 127–135. ISSN 1434-8411. DOI: 10.1016/j.aeue.2017.08.010.

[16] ADHIKARY, A., S. SEN and K. BISWAS. Practical realization of tunable fractional order parallel resonator and fractional order filters. *IEEE Transactions on Circuits and Systems I: Regular Papers*. 2016, vol. 63, no. 8, pp. 1142–1151. ISSN 1549-8328. DOI: 10.1109/TCSI.2016.2568262.

[17] BERTSIAS, P., F. KHATEB, D. KUBANEK, F. A. KHANDAY and C. PSYCHALINOS. Capacitorless digitally programmable fractional-order filters. *AEU-International Journal of Electronics and Communications*. 2017, vol. 78, iss. 4, pp. 228–237. ISSN 1434-8411. DOI: 10.1016/j.aeue.2017.04.030.

[18] TSIRIMOKOU, G., C. PSYCHALINOS and A. S. ELWAKIL. Fractional-order electronically controlled generalized filters. *International Journal of Circuit Theory and Applications*. 2017, vol. 45, no. 5, pp. 595–612. ISSN 1097-007X. DOI: 10.1002/cta.2250.

[19] MAUNDY, B., A. ELWAKIL and S. GIFT. On a multivibrator that employs a fractional capacitor. *Analog Integrated Circuits and Signal Processing*. 2010, vol. 62, no. 1, pp. 99–103. ISSN 0925-1030. DOI: 10.1007/s10470-009-9329-3.

[20] TSENG, C.-C. Design of FIR and IIR fractional order Simpson digital integrators. *Signal Processing*. 2007, vol. 87, no. 5, pp. 1045–1057. ISSN 0165-1684. DOI: 10.1016/j.sigpro.2006.09.006.

[21] KRISHNA, B. T. Studies on fractional order differentiators and integrators: A survey. *Signal Processing*. 2011, vol. 91, no. 3, pp. 386–426. ISSN 0165-1684. DOI: 10.1016/j.sigpro.2010.06.022.

[22] AGAMBAYEV, A., S. PATOLE, H. BAGCI and K. N. SALAMA. Tunable fractional-order capacitor using layered ferroelectric polymers. *AIP Advances*. 2017, vol. 7, no. 9, pp. 1–8. ISSN 2158-3226. DOI: 10.1063/1.4991659.

[23] KRISHNA, B. T. and K. V. V. S. REDDY. Active and passive realization of fractance device of order 1/2. *Active and passive electronic components*. 2008, vol. 2008, no. 369421, pp. 1–6. ISSN 1563-5031. DOI: 10.1155/2008/369421.

[24] OLDHAM, K. and J. SPANIER. *The fractional calculus theory and applications of differentiation and integration to arbitrary order*. New York: Elsevier, 1974. ISBN 0-12-525550-0.

[25] NAKAGAWA, M. and K. SORIMACHI. Basic characteristics of a fractance device. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*. 1992, vol. 75, no. 12, pp. 1814–1819. ISSN 0916-8508.

[26] SUGI, M., Y. HIRANO, Y. F. MIURA and K. SAITO. Simulation of fractal immittance by analog circuits: an approach to the optimized circuits. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*. 1999, vol. 82, no. 8, pp. 1627–1635. ISSN 0916-8508.

[27] ROY, S. D. On the realization of a constant-argument immittance or fractional operator. *IEEE Transactions on Circuit Theory*. 1967, vol. 14, no. 3, pp. 264–274. ISSN 0018-9324. DOI: 10.1109/TCT.1967.1082706.

[28] CARLSON, G. and C. HALIJAK. Approximation of fractional capacitors (1/s)^(1/n) by a regular Newton process. *IEEE Transactions on Circuit Theory*. 1964, vol. 11, no. 2, pp. 210–213. ISSN 0018-9324. DOI: 10.1109/TCT.1964.1082270.

[29] STEIGLITZ, K. An RC Impedance Approximant to s^{-1/2}. *IEEE Transactions on Circuit Theory*. 1964, vol. 11, no. 1, pp. 160–161. ISSN 0018-9324. DOI: 10.1109/TCT.1964.1082252.

[30] SIEROCIUK, D., I. PODLUBNY and I. PETRAS. Experimental evidence of variable-order behavior of ladders and nested ladders. *IEEE Transactions on Control Systems Technology*. 2013, vol. 21, no. 2, pp. 459–466. ISSN 1063-6536. DOI: 10.1109/TCST.2012.2185932.

[31] SUGI, M., Y. HIRANO, Y. F. MIURA and K. SAITO. Frequency behavior of self-similar ladder circuits. *Colloids and Surfaces A: Physicochemical and Engineering Aspects*. 2002, vol. 198, iss. 1, pp. 683–688. ISSN 0927-7757. DOI: 10.1016/S0927-7757(01)00988-8.

[32] OLDHAM, K. B. and C. G. ZOSKI. Analogue instrumentation for processing polarographic data. *Journal of electroanalytical chemistry and interfacial electrochemistry*. 1983, vol. 157, no. 1, pp. 27–51. ISSN 1572-6657. DOI: 10.1016/S0022-0728(83)80374-X.

[33] MONDAL, D. and K. BISWAS. Packaging of single-component fractional order element. *IEEE Transactions on Device and Materials Reliability*. 2013, vol. 13, no. 1, pp. 73–80. ISSN 1530-4388. DOI: 10.1109/TDMR.2012.2212020.

[34] ADHIKARY, A., M. KHANRA, S. SEN and K. BISWAS. Realization of a carbon nanotube based electrochemical fractor. In: *2015 IEEE International Symposium on Circuits and Systems (ISCAS)*. Lisbon: IEEE, 2015, pp. 2329–2332. ISBN 978-1-4799-8391-9. DOI: 10.1109/ISCAS.2015.7169150.

[35] SALAMA, K. N. and A. M. SOLIMAN. CMOS operational transresistance amplifier for analog signal processing. *Microelectronics Journal*. 1999, vol. 30, no. 3, pp. 235–245. ISSN 0026-2692. DOI: 10.1016/S0026-2692(98)00112-8.

[36] KAFRAWY, A. K. and A. M. SOLIMAN. A modified CMOS differential operational transresistance amplifier (OTRA). *AEU-International Journal of Electronics and Communications*. 2009, vol. 63, iss. 12, pp. 1067–1071. ISSN 1434-8411. DOI: 10.1016/j.aeue.2008.08.003.

[37] GHOSH, M., S. K. PAUL, R. K. RANJAN and A. RANJAN. Third order universal filter using single operational transresistance amplifier. *Journal of Engineering*. 2013, vol. 2013, no. 317296, pp. 1–6. ISSN 2314-4904. DOI: 10.1155/2013/317296.

[38] PANDEY, R., N. PANDEY, G. KOMANAPALLI and R. ANURAG. OTRA based voltage mode third order quadrature oscillator. *ISRN Electronics*. 2014, vol. 2014, no. 126471, pp. 1–5. ISSN 2090-8679. DOI: 10.1155/2014/126471.

[39] LO, Y.-K., H.-C. CHIEN and H.-J. CHIU. Switch-controllable OTRA-based bistable multivibrators. *IET circuits, devices & systems*. 2008, vol. 2, no. 4, pp. 373–382. ISSN 1751-858X. DOI: 10.1049/ietcds:20080011.

[40] PANDEY, R., N. PANDEY, S. K. PAUL, A. SINGH, B. SRIRAM and K. TRIVEDI. Novel grounded inductance simulator using single OTRA. *International Journal of Circuit Theory and Applications*. 2014, vol. 42, no. 10, pp. 1069–1079. ISSN 1097-007X. DOI: 10.1002/cta.1905.

[41] ANSARI, A. Q., G. KAUR and M. S. HASHMI. Current Differencing Buffered Amplifier (CDBA) based Current Mode Universal Fractional Order Filter. In: *National Conference on Advances in*

Microelectronics, Instrumentation and Communication (MICOM 2015). Pilani: BITS, 2015, pp. 1–4. ISBN 978-93-86256-11-9.

[42] SOLTAN, A., A. G. RADWAN and A. M. SOLIMAN. Fractional order Sallen-Key and KHN filters: stability and poles allocation. *Circuits, Systems, and Signal Processing*. 2015, vol. 34, no. 5, pp. 1461–1480. ISSN 0278-081X. DOI: 10.1007/s00034-014-9925-z.

[43] RADWAN, A. G., A. M. SOLIMAN, A. S. ELWAKIL and A. SEDEEK. On the stability of linear systems with fractional-order elements. *Chaos, Solitons & Fractals*. 2009, vol. 40, no. 5, pp. 2317–2328. ISSN 0960-0779. DOI: 10.1016/j.chaos.2007.10.033.

[44] TANWAR, R. and S. KUMAR. Analysis & design of fractance based fractional order filter. *International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering*. 2013, vol. 1, no. 3, pp. 113–122. ISSN 2321-2004.

[45] SHAH, Z. M., M. Y. KATHJOO, F. A. KHANDAY, K. BISWAS and C. PSYCHALINOS. A survey of single and multi-component Fractional-Order Elements (FOEs) and their applications. *Microelectronics Journal*. 2019, vol. 84, iss. 1, pp. 9–25. ISSN 0026-2692. DOI: 10.1016/j.mejo.2018.12.010.

About Authors

Gagandeep KAUR received her B.Tech and M.E. degrees in 2002 and 2011 from Punjab Technical University, Jalandhar (Punjab) and Delhi College of Engineering (DTU), University of Delhi, Delhi respectively. She has been teaching UG students in engineering since last nine years in institutes of repute in Delhi/NCR. Currently, she is working as an Assistant Professor in the Department of Electronics and Communication Engineering at Guru Tegh Bahadur Institute of Technology, Delhi. Her research interests include Linear Integrated Circuits, Analog Signal Processing and Fractional order Circuits.

Abdul Quaiyum ANSARI received his B.Sc. Eng. (Hons.) in Electrical Engineering (low current) from Aligarh Muslim University 1984, M.Tech. from Indian Institute of Technology Delhi in Integrated Electronics and Circuits in 1991 and Ph.D. in the field of Hierarchical Fuzzy Control for Industrial Automation from Jamia Millia Islamia (JMI), New Delhi in 2000. He joined as Lecturer in Jamia Millia Islamia in 1984. He was Professor and Head of the Department at Jamia Hamdard (Hamdard University) from 2001 to 2004. He was Head of the Department at Jamia Millia Islamia from 2008 to 2011. He was Dean in Faculty of Management Studies and Information Technology, Jamia Hamdard from 2002 to 2004. (On deputation from JMI) He is professor of Electrical Engineering department in Jamia Millia Islamia since 2000. He is Honorary Executive Secretary Indian Society for Technical Education (ISTE) from April 2005 to July 2005, Executive Council Member National Executive Council, Indian Society for Technical Education (ISTE) from January 2003 to December 2006, Chairman IEEE-Computational Intelligence Society, Delhi Chapter January 2011 till date, Chairman Standing committee on membership development, IEEE Delhi Section from January 2014 till date, EC Member IEEE Delhi Section from January 2009 till date. His research contributions are in the areas of Mobile Adhoc Networks, Multimodal Biometrics, Networks-on-Chip, Fuzzy Logic and its Variants and fractional order circuits.

Mohammad Shabi HASHMI attended Prifysgol Caerdydd (Cardiff University) for a Ph.D. degree in Electronics Engineering, Technische Universit Darmstadt for an M.Sc. degree in Information and Communication Engineering, and Aligarh Muslim University for a B.Tech. degree in Electrical Engineering. He is currently a faculty member of Electronics and Communications Engineering at Indraprastha Institute of Information Technology Delhi, New Delhi, India. Earlier he held research and engineering positions at the University of Calgary in Canada, Cardiff University in the UK, Philips Technology Centre in Nuremberg, and Thales Engineering Design Centre in Berlin. He works in the broad area of Electronics System Design with particular emphasis on RF, Mixed Signal Electronics and Fractional order Circuits and Systems. His research has been funded by Nokia Networks, Focus Microwaves, and Agilent Technologies. He has authored or co-authored more than 60 peer reviewed journal and conference papers, 1 book, and 3 US Patents (2 pending). His major awards and achievements include DAAD-Siemens Scholarship, Nokia Doctoral Fellowship, UK Govt's Dorothy Hodgkin Award, Alberta Innovates Fellowship, third prize in the IEEE-MTT originality and creativity competition (2008), ARFTG Microwave Measurement Fellowship (2008), and Young Researcher Grant from INMMiC (2008).

Hybrid Control Approach using NLMS and PLMS Algorithms for Grid Connected SPV System

M. Rizwan, and Priyanka Chaudhary

***Abstract*— The control of grid tied solar PV (SPV) system using combination of normalized least mean square (NLMS) algorithm with proportionate normalized least mean square (PNLMS) algorithm is presented in this paper. A PI controller is implemented to regulate the DC link voltage. INC (increment conductance) based maximum power point tracking (MPPT) controller to enhance the maximum power operation of SPV array. The developed system support the three phase AC grid by supplying power and to connected loads with improved power quality. The given inverter control approach takes out the fundamental load current components for reference current generation and to synchronize the SPV system with the grid with very fast response under rapid change scenario. The proposed system includes a SPV array, inverter, filter to take out the ripple and DC components and different types of loads. System provides harmonics in grid currents and voltage within limits as per IEEE-519 and IEEE-1547 standards.**

***Index Terms*— Solar Photovoltaic (SPV) system, inverter, maximum power point (MPPT), increment conductance, normalized least mean square (NLMS) algorithm, proportionate normalized least mean square (PNLMS)**

I. Introduction

SPV systems are most widely used renewable energy sources (RES) now a days which possess numerous advantages. These advantages includes locally available generation, high efficiency, less maintenance and minimal environmental impact. Grid integration of SPV power generation systems at the grid side enhance the generation and also create adverse effects such as voltage limit violation, frequency fluctuation, grid instability etc. [1-3]. Grid connected SPV array must follow the codes and regulations of grid defined by the IEEE 1547, IEC 61727 and VDE-AR-N4105. Inverter control in a grid integrated mode of SPV power generation system plays very important role to synchronizing it with grid. Numerous control approaches for grid synchronization and integration of SPV power generation system are widely available in literature [4-7].

Non-linear nature of loads is main source of power quality (PQ) issues which leads too reactive power consumption and distortion of grid currents. All these problems need to be mitigate as causes degradation of power factor and force the system to operate with low power factor [8]. Varma et. al. [20] have developed a PV-STATCOM system for enhancing the power transfer capabilities to transmit active power. Hamid et. al. [21] reported an approach to mitigate the current harmonics from a SPV system by placing a power conditioner unit in parallel with the power generation system which operates in feed-forward mode to compensate the output current distortion.

In this paper, combination of NLMS algorithm with PNLMS algorithm is used to control grid tied PV inverter. A PI controller is used to regulate the DC link voltage and maintain it to reference value. INC (increment conductance) based maximum power point tracking (MPPT) controller to enhance the maximum power operation of SPV array. The modelling and simulations of the given system are performed using MATLAB/Simulink platform.

M. Rizwan is with Delhi Technological University, Delhi, India. He is working as associate professor in the Department of Electrical Engineering, DTU, Delhi, India. (e-mail: rizwan@dce.ac.in).

Priyanka Chaudhary is with Delhi Technological University, Delhi, India. She is working as research scholar in the Department of Electrical Engineering, DTU, Delhi, India. (e-mail: priyankach.iilm@gmail.com).

II. System Description

The developed system consists of a 10 kW SPV array, voltage source inverter (VSC), various types of loads such as linear and nonlinear loads, ripple filter, three phase AC grid as shown in Fig.1. Ripple filter which is the combination of R and C used to remove the excessive fluctuations from the inverter output.

A. Solar PV Array

A SPV array combines several PV modules which are again formed by series and parallel interconnection of various PV cells. SPV array is designed as per the power requirement of applications, PV modules are connected in series configuration if high voltage is required and for high current application parallel configuration is used. The output current equation can be written as [8]:

$$I_{PV} = I_{sc}N_p - N_s I_o \left[\exp\left\{\frac{q(V_{PV}+I_{PV}Rs)}{N_s AkT}\right\} - 1\right] V_{PV} + \frac{I_{PV}Rs}{Rp} \quad (1)$$

Where reverse saturation current I_{rs} is:

$$I_{rs} = I_{scref} + [\exp\left(\frac{qVoc}{NskAT}\right) - 1] \quad (2)$$

The I_o saturation current as a function of temperature can be written as:

$$I_o = I_{rs}\left[\frac{\left(\frac{T}{T_{ref}}\right)^3 e^{qCg}}{Ak} * \left(\frac{1}{T_{ref}} - \frac{1}{T}\right)\right] \quad (3)$$

Where, k denotes the Boltzmann constant (1.38 x 10^{-23} J K^{-1}), q the electronic charge (1.602 x 10^{-19} C), T cell temperature (K); A diode ideality factor, R_s the series

ISBN: 978-988-14048-6-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

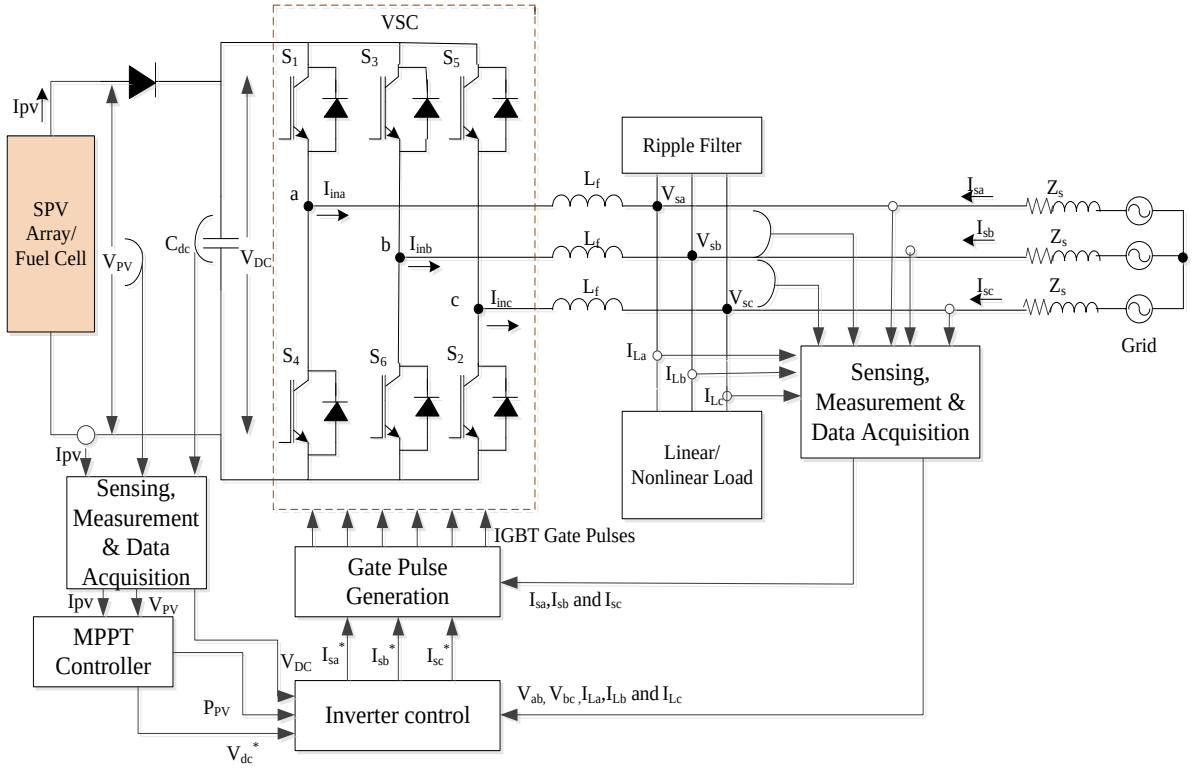


Fig. 1. System description

resistance (Ω) and R_p is the shunt resistance (Ω). N_S and N_p are number of cells connected in series and parallel respectively.

B. Inverter Design

Inverter consists of three arms and each arm contains two IGBT with antiparallel diodes and designed for 415 V, 25 kW at 0.8 p.f. lagging load. A combination of series resistance and capacitance is used as ripple filter and design values are given in appendix []. The appropriate design of interfacing inductor for inverter plays very important for tracking the reference current.

III. Control Design

A. Maximum Power Point Tracking

A maximum power point tracking (MPPT) technique is required to extract the maximum power from Solar PV (SPV) array. Various MPPT techniques are presented in the literature [9-10] which includes perturb and observe (P&O), incremental conductance (INC), constant voltage, open circuit voltage, short circuit current, extremum seeking control and techniques based on artificial neural networks, fuzzy logic, genetic algorithms. INC technique has been implemented in present work and works well than other available techniques [11]. This method utilizes the fact that the slope of PV array output curve is always equals to zero at maximum power point (MPP). This curve has positive slope for values less than MPP and negative for which has values more than MPP.

B. Inverter Control

Inverter control plays very important role for design and development of grid integrated solar PV systems. The knowledge of phase angle and frequency of grid side voltage is necessary to for inverter control. The reference sinusoidal source currents for generating inverter gate pulses obtained by using present control algorithm. The implemented control algorithm combines normalized least mean square (NLMS) algorithm with proportionate normalized least mean square (PNLMS) algorithm and a PI controller to regulate the DC link voltage. This results in a more stability with high convergence rate.

Amplitude of terminal voltage and unit templates:

Measured grid side line voltages (v_{ab}, v_{bc}, v_{ca}) at point of common coupling (PCC) is used to find grid phase voltages (V_{sa}, V_{sb} andV_{sc}) calculated as:

$$V_{sa} = \frac{2v_{ab}+v_{bc}}{3}, V_{sb} = \frac{-v_{ab}+v_{bc}}{3}, V_{sc} = \frac{-v_{ab}-2v_{bc}}{3} \quad (4)$$

The terminal voltage amplitude (V_t) at PCC can be computed as:

$$V_t = \sqrt{\left[\frac{2\left({V_{sa}}^2 + {V_{sb}}^2 + {V_{sc}}^2\right)}{3}\right]} \quad (5)$$

The in-phase unit templates ofv_a, v_b and v_c are obtained as:

$$u_{pa} = \frac{V_{sa}}{V_t}, u_{pb} = \frac{V_{sb}}{V_t}, u_{pc} = \frac{V_{sc}}{V_t} \quad (6)$$

Quadrature unit templates can be obtained from in phase unit templates as following:

$$u_{qa} = \frac{u_{pc}}{\sqrt{3}} - \frac{u_{pb}}{\sqrt{3}}, u_{qb} = \frac{\sqrt{3}u_{pa}}{2} + \frac{u_{pb}-u_{pc}}{2\sqrt{3}}, u_{qc} = \frac{u_{pb}-u_{pc}}{2\sqrt{3}} - \frac{\sqrt{3}u_{pa}}{2} \quad (7)$$

Reference grid currents generation:

Voltage at dc link (V_{dc}) is measured and compared with the reference dc link voltage (V_{dc}^*) to calculate the active loss component.

$$E_{dc}(k) = V_{dc}^*(k) - V_{dc}(k) \quad (8)$$

ISBN: 978-988-14048-6-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

PI controller is used to minimal this difference (E_{dc}) which is further being used to get active current component(w_{pdc}). The controller's output at $(\mathrm{k}+1)^{\mathrm{th}}$ instant:

$$w_{\mathrm{pd}}(k+1) = w_{\mathrm{pd}}(k) + k_{pd}\{V_{dce}(k+1) - V_{dce}(k)\} + k_{id}\, E_{dc}(k+1) \quad (9)$$

Error V_{te} (difference between sensed and reference terminal voltage $V_t$and V_t^* respectively) at point of common coupling passed through a PI controller to compute reactive loss component which needed to maintain the constant terminal voltage.

$$E_{te}(k) = \mathrm{V_t^*(k)} - \mathrm{V_t}(k) \quad (10)$$

The controller output at $(\mathrm{k}+1)^{\mathrm{th}}$ is:

$$w_{\mathrm{qt}}(k+1) = w_{\mathrm{qt}}(k) + k_{pt}\{V_{te}(k+1) - V_{te}(k)\} + k_{it}E_{te}(k+1) \quad (11)$$

Where, k_{pt} and k_{it} are proportional and integral gains respectively. Fast dynamic response can be achieved with a feed forward weight which is a function of solar power (P_{PV}) andV_t.

$$\mathrm{w}_{PV}(k) = \frac{2\mathrm{P}_{PV}(k)}{3\mathrm{V}_t} \quad (12)$$

Fundamental active and reactive components of load current extraction done by taking load currents ($i_{\mathrm{La}}, i_{\mathrm{Lb}}, i_{\mathrm{Lc}}$). The error for the system is considered as:

$$e_{Aj}(k) = \mathrm{i_{Lj}}(k) - u_{pj}(k) * w_{Pj}(k) \quad (13)$$

In above equation n shows the phases a, b and c, w_{pj} depicts the fundamental active load current component of phase 'j'.

$$w_{\mathrm{Pj}}(k+1) = w_{\mathrm{Pj}}(k) + \mu * u_{pj}(k) * e_{Aj}(k) * [g_{Aj}(k) + h_{Aj}(k)] \quad (14)$$

Here μ is adaption constant and values lie between 0 and 2. NLMS algorithm contribute as$g_{Aj}(k)$:

$$g_{Aj}(k) = \frac{(1-\alpha)}{2*u_{pj}(k)*u_{pj}^{\mathrm{T}}(\mathrm{k})+\varepsilon} \quad (15)$$

PNLMS algorithm contribute as$h_{Aj}(k)$:

$$h_{Aj}(k) = \frac{(1+\alpha)*p_{Aj}(k-1)}{2*u_{pj}(k)*p_{Aj}(k-1)*u_{pj}^{\mathrm{T}}(\mathrm{k})+\varepsilon} \quad (16)$$

Where, 0<α<-0.5 and regulation constant 0.001<ε<0.01 for finite output. $p_{Aj}(k)$ [12] can be written as:

$$p_{Aj}(k) = \frac{|w_{\mathrm{Pj}}(k)|}{\sum_{i=0}^{M-1}|w_{\mathrm{Pj}}(k)|+\delta} + \gamma \quad (17)$$

Where, δ and γ are considered as positive small constant for finite $p_{Aj}(k)$.

Weighted values for fundamental active current components for all phases:

$$\left.\begin{aligned} w_{\mathrm{Pa}}(k+1) &= w_{\mathrm{Pa}}(k) + \mu * u_{pa}(k) * e_{Aa}(k) * [g_{Aa}(k) + h_{Aa}(k)] \\ w_{\mathrm{Pb}}(k+1) &= w_{\mathrm{Pb}}(k) + \mu * u_{pb}(k) * e_{Ab}(k) * [g_{Ab}(k) + h_{Ab}(k)] \\ w_{Pa}(k+1) &= w_{Pc}(k) + \mu * u_{pc}(k) * e_{Ac}(k) * [g_{Ac}(k) + h_{Ac}(k)] \end{aligned}\right\} \quad (18)$$

The arithmetic mean active component of load currents (w_{Lp}) is calculated as below:

$$w_{Lp} = \frac{(w_{\mathrm{Pa}}+w_{\mathrm{Pb}}+w_{\mathrm{Pc}})}{3}. \quad (19)$$

Similarly, weighted values for mean reactive component of load currents (w_{qj}) is obtained as:

$$w_{Lq} = \frac{(w_{\mathrm{qa}}+w_{\mathrm{qb}}+w_{qc})}{3}. \quad (20)$$

Net fundamental active load current component is given as:

$$I_{\mathrm{Lp}} = w_{\mathrm{pd}} + w_{\mathrm{Lp}} - w_{\mathrm{PV}} \quad (21)$$

Net fundamental reactive current component of load current:

$$I_{\mathrm{Lq}} = w_{\mathrm{qt}} + w_{\mathrm{Lq}} \quad (22)$$

The reference grid currents for active and reactive components are given in equation 23 and 24 respectively:

$$i_{\mathrm{psa}} = I_{\mathrm{Lp}} * u_{pa}, i_{\mathrm{psb}} = I_{\mathrm{Lp}} * u_{pb}, i_{\mathrm{psc}} = I_{\mathrm{Lp}} * u_{pc} \quad (23)$$

$$i_{\mathrm{qsa}} = I_{\mathrm{Lq}} * u_{qa}, i_{\mathrm{qsb}} = I_{\mathrm{Lq}} * u_{qb}, i_{\mathrm{qsc}} = I_{\mathrm{Lq}} * u_{qc} \quad (24)$$

Net reference grid currents ($i_{sa}^*, i_{sb}^*, i_{sc}^*$) is given as:

$$i_{sa}^* = i_{\mathrm{psa}} + i_{\mathrm{qsa}},\ i_{sb}^* = i_{\mathrm{psb}} + i_{\mathrm{qsb}},\ i_{sc}^* = i_{\mathrm{psc}} + i_{\mathrm{qsc}} \quad (25)$$

Gate pulses for inverter switches are generated by comparing reference ($i_{sa}^*, i_{sb}^*, i_{sc}^*$) and actual ($i_{\mathrm{sa}}, i_{\mathrm{sb}}, i_{\mathrm{sc}}$) grid currents.

IV. Results

The performance of developed system has been tested for various cases such as STC conditions, various irradiance levels with different kind of loads such as linear, nonlinear, balanced and unbalanced. Proposed model has been validated through analysis of various parameters such as grid voltage ($\mathrm{V_s}$), DC bus voltage ($\mathrm{V_{dc}}$), AC grid currents, ($\mathrm{I_s}$), load currents, ($\mathrm{I_L}$) the real power from and to the grid (Pg), the reactive power ($\mathrm{Q_g}$), SPV current ($\mathrm{I_{pv}}$), SPV voltage ($\mathrm{V_{pv}}$), SPV power ($\mathrm{P_{pv}}$), inverter current ($\mathrm{I_{inv}}$) and AC terminal voltage ($\mathrm{V_t}$). Fig. 2 shows results of the system for dynamic linear load scenario.

Fig. 3 presents dynamic operation of the system under unbalanced nonlinear loads for zero voltage regulation (ZVR) mode. At 1.1 s phase 'c' of load is removed from supply, even under this situation, source currents (i_s) maintained sinusoidal. The proposed control helps $\mathrm{V_t}$ and $\mathrm{V_{dc}}$ are maintained at 415 V and 700 V, respectively without fluctuations. Performance of the given system under variable solar irradiance is shown in Fig. 4. Fig. 5 shows the THDs at PCC for load current phase 'a' and grid current respectively and it can be observe that THD within the limits according to IEEE-519 and IEEE-1547 standards [13].

ISBN: 978-988-14048-6-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

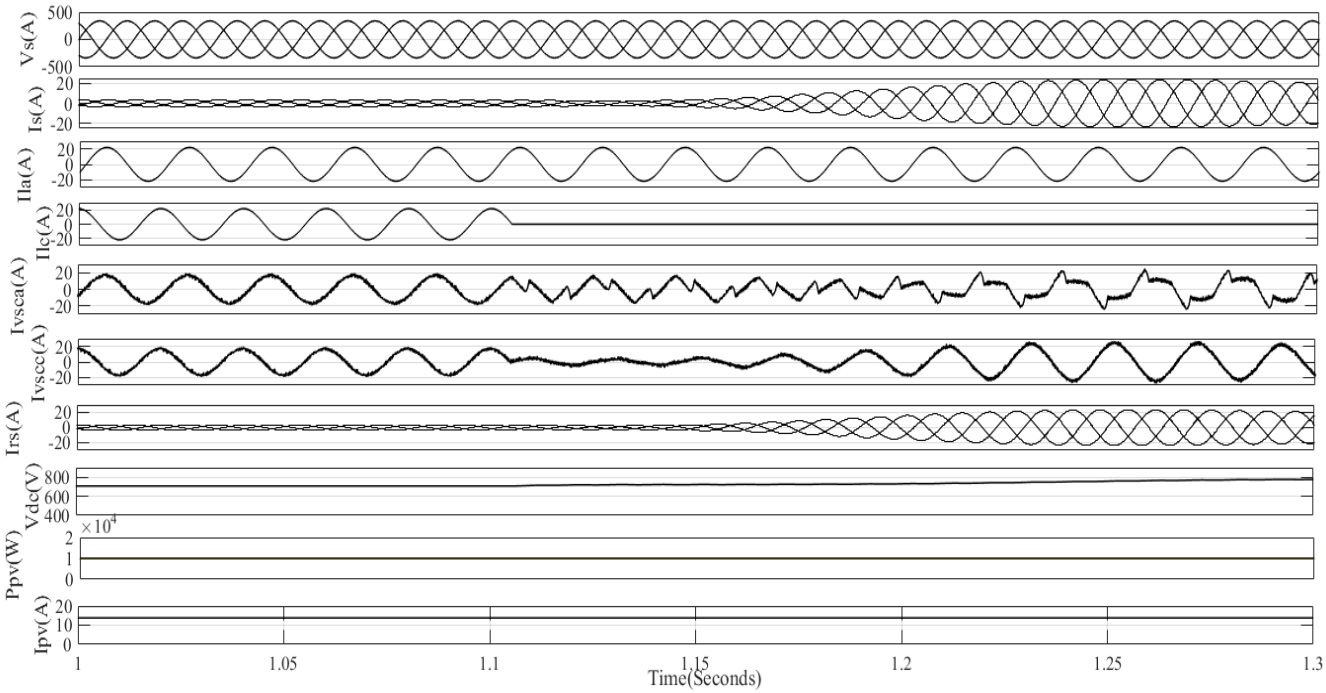


Fig. 2. Performance under linear dynamic load.

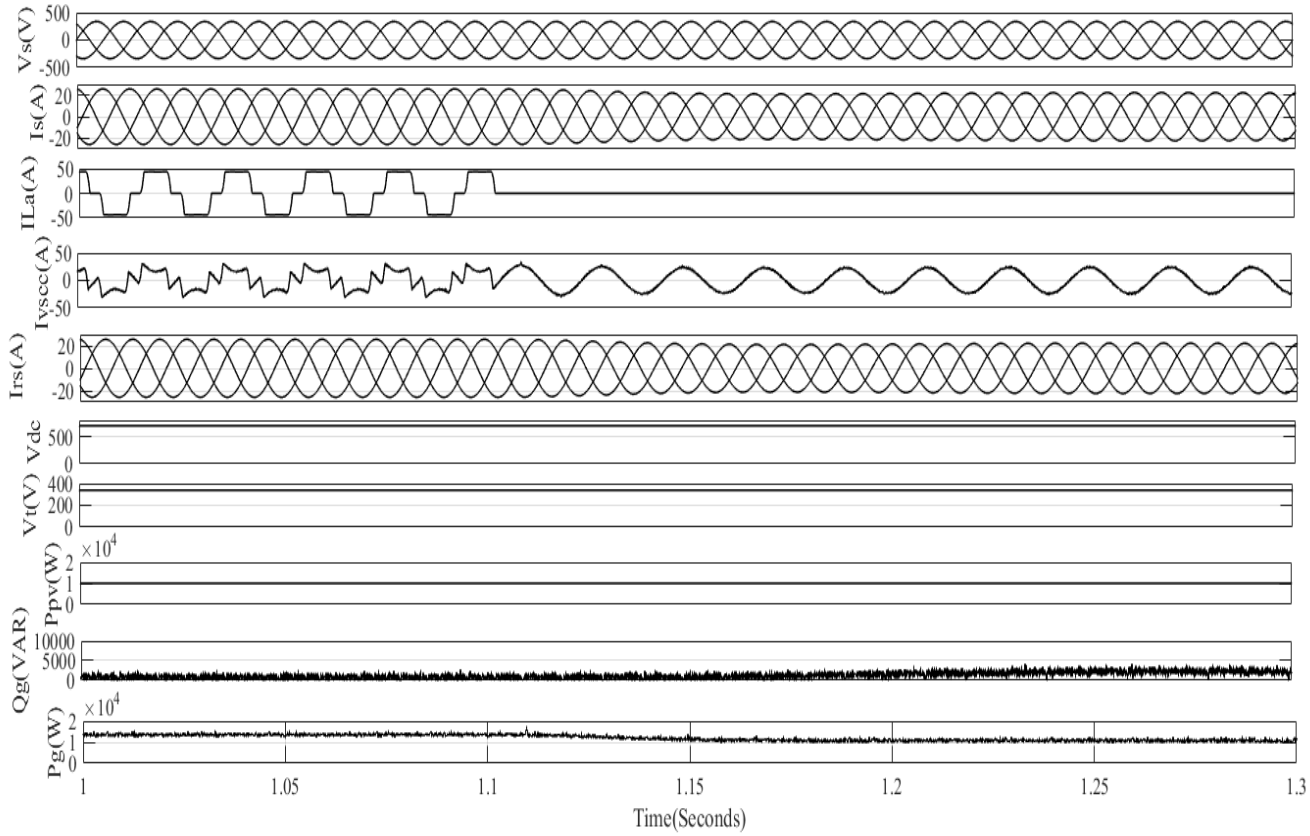


Fig. 3. Performance under non-linear dynamic load.

ISBN: 978-988-14048-6-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

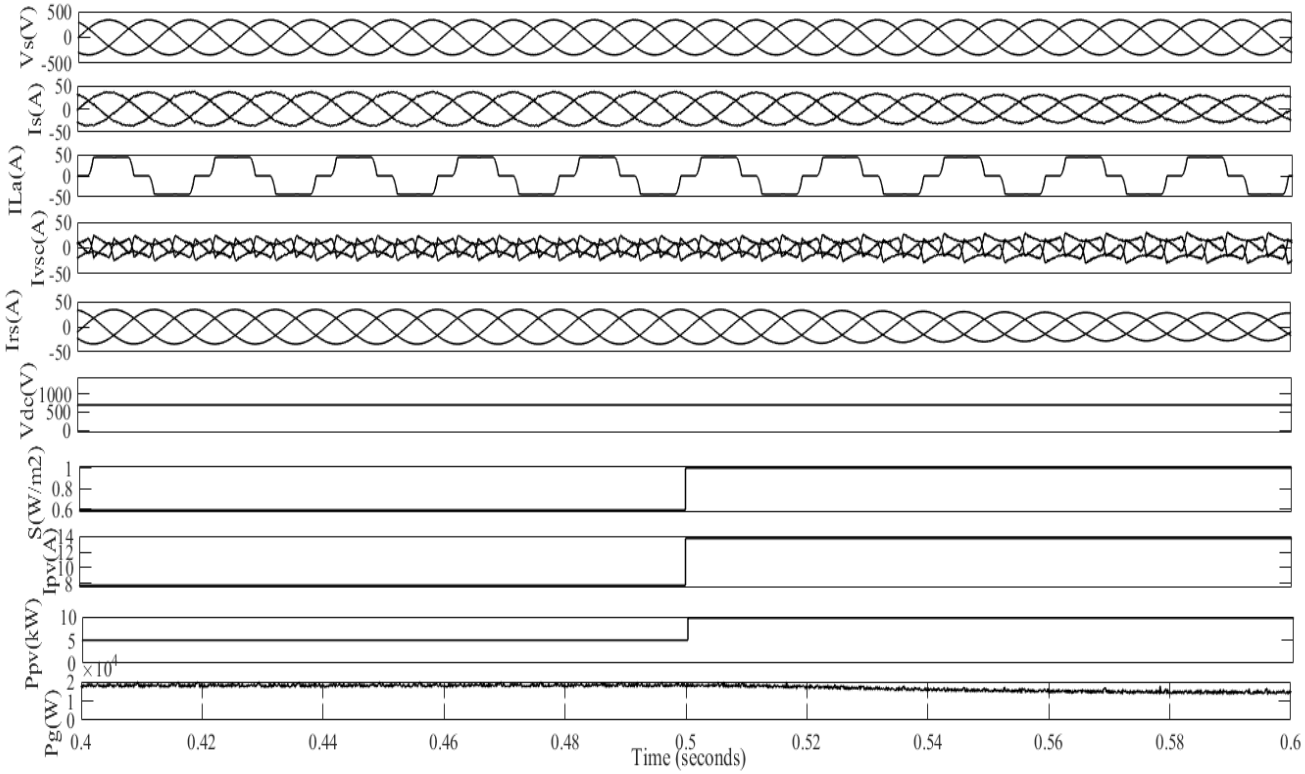


Fig. 4. Performance under variable solar irradiance.

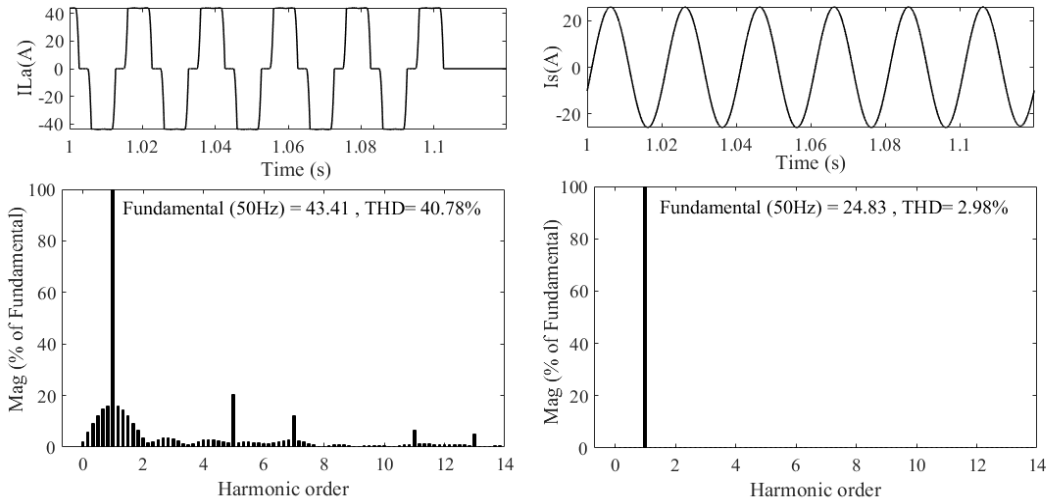


Fig. 5. THD spectrum for load current and grid current.

V. Conclusion

The proposed system is developed in MATLAB/Simulink platform. Performance of the proposed control approach is analyzed for linear and nonlinear loads with steady state and dynamic states. Load unbalancing and solar irradiance variation to validate the results of the control algorithm and satisfactory performance has been achieved. System provides fast response and operates unity power factor (UPF) mode gives reactive power compensation, load balancing, maximum power extraction and harmonic reduction. Grid currents harmonics and voltage fluctuations are within limits as per IEEE-519 and IEEE-1547 standards.

Appendix

The developed system design specifications.

AC Grid specifications= 3 ϕ, 415 V, 50 Hz, SPV array voltage (V_{MPP}) = 700V, SPV array current (I_{MPP}) = 13.5A, SPV array power (P_{MPP}) = 10kW, Interfacing inductor (L_{fa}= L_{fb}= L_{fc}) = 2.6 mH, DC bus capacitor (C_{DC}) = 10mf, DC bus voltage (V_{DC}) = 700V, Ripple Filter (R_f) and (C_f) = 5 Ω, 10 μF.

ISBN: 978-988-14048-6-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

REFERENCES

[1] S. J. Steffel, P. R. Caroselli, A. M. Dinkel, J. Q. Liu, R. N. Sackey, and N. R. Vadhar, "Integrating solar generation on the electric distribution grid", IEEE Trans. Smart Grid, vol. 3 (2), 2012, pp. 878-886.

[2] Y. W. Li, J. He, "Distribution system harmonic compensation methods: an overview of DG-interfacing inverters", IEEE Trans. Ind. Elect., vol. 8, no. 4, Dec. 2014, pp. 18-31.

[3] P. Chaudhary and M. Rizwan M, "Voltage regulation mitigation techniques in distribution system with high PV penetration: A review", Renewable and Sustainable Energy Reviews, vol. 82, no. 3, 2018, pp. 3279-3287.

[4] N. Beniwal, I. Hussain and B. Singh, "Implementation of DSTATCOM with i-PNLMS based control algorithm under abnormal grid conditions"' 7th India International Conference on Power Electronics (IICPE), Patiala 2016, pp. 1-5.

[5] R. K. Agarwal, I. Hussain and B. Singh, "Three-phase single-stage grid tied solar PV ECS using PLL-less fast CTF control technique", IET Power Electronics, vol. 10, no. 2, 2017, pp. 178-188.

[6] S. Adhikari and Li. Fangxing, "Coordinated V-f and P-Q control of solar photovoltaic generators with MPPT and battery storage in microgrids", IEEE Transactions on Smart Grid, vol. 5, no. 3, 2014, pp. 1270–1281.

[7] P. Chaudhary and M. Rizwan, "A Predictive Current Control for Solar PV Fed VSI in Distribution System," 17th IEEE International Conference on Environment and Electrical Engineering (IEEE EEEIC17), June 6-9, 2017, Milan, Italy.

[8] J. A. Gow and C. D. Manning, "Development of a photovoltaic array model for use in power-electronics simulation studies," IEEE Proceedings on Electric Power Applications, vol. 146, no. 2, 1999, pp. 193–200.

[9] A. R. Reisi, M. H. Moradi and S. Jamsab, "Classification and comparison of maximum power point tracking techniques for photovoltaic system: A review", Renewable and Sustainable Energy Reviews, vol. 19, 2013, pp. 433-443.

[10] Y. C. Kuo, T. J. Liang and J. F. Chen, "Novel maximum-power point tracking controller for photovoltaic energy conversion system", IEEE Transaction on Industrial Electronics, vol. 48, no. 3, 2001, pp. 594–601.

[11] M. Rizwan, P. Chaudhary and T. Ahmad, "Performance Analysis of Maximum Power Point Tracking Techniques for Photovoltaic Systems", Advanced Science Letters, vol. 20, 2014, pp. 1231-1247.

[12] A. Maleki and K. Nayebi, "A new efficient PNLMS based algorithm for adaptive line echo cancellation", Seventh International Symposium on Signal Processing and Its Applications, vol. 2, 2003, pp. 555-558.

[13] IEEE Recommended Practices and Requirements for Harmonics Control in Electric Power Systems, IEEE Standard 519, 1992.

ISBN: 978-988-14048-6-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

INVESTIGATION OF LOCAL OPENING-SLIDING RELATIONSHIP IN THE VICINITY OF DEFORMED BAR IN CONCRETE BY USING DIC TECHNIQUE

A. OKEIL[*], R. UPADHYAY[†], K. MATSUMOTO[††] AND K. NAGAI[†††]

[*] The University of Tokyo
Tokyo, Japan
E-mail: okeil@iis.u-tokyo.ac.jp

[†] Delhi Technological University
Delhi, India
E-mail: rheaupadhyay04@gmail.com

[††] The University of Tokyo
Tokyo, Japan
E-mail: km312@iis.u-tokyo.ac.jp

[†††] The University of Tokyo
Tokyo, Japan
E-mail: nagai325@iis.u-tokyo.ac.jp

Key words: Local bond, Crack propagation, Aggregates, Deformed bars, Image processing

Abstract: One of the important factors affecting the behaviour of a reinforced concrete (RC) is the bond between concrete and the reinforcement bars. To model the bond between concrete and the reinforcement in the numerical simulation of a RC, accurately modelling for the reinforcing bar with its ribs in 3D shape is required. So to simulate the bond behaviour between concrete and steel rebar, the contact surface between them is represented by normal and shear behaviours. The characteristics of these springs are modified based on numerical simulations to get best-matched results to the experimental results. Therefore, in this study, the opening-sliding relationships between a deformed bar and surrounding matrix are investigated experimentally by using digital images correlation (DIC) technique. A standard axial tension test was applied to a deformed steel bar that was embedded in concrete and cement paste specimen. The specimen had a trapezoidal "window," which allows full continues observation and measurements of concrete or cement paste displacements by recording high-resolution digital video that was later analysed by DIC software. The half of the steel bar's cross-section was removed to see directly the ribs of the rebar and thereby capture the local deformations between these ribs. For that purpose 15 points were chosen, at the tips of the ribs and also between the ribs of the deformed bar, to investigate the relative displacements between the deformed bar and surrounding matrix in X-direction and Y-direction. Based on the results of the experiments, some relationships between opening and sliding in case of cement paste or concrete were established. It was found that the opening between the concrete and the rebar at the tip of the rib is higher than between the ribs. Also, the sliding between the concrete and the rebar at the tip of the rib is higher than between the ribs.

1 INTRODUCTION

Concrete-steel composite is widely used as a structural material. The bond between concrete and steel is necessary not only to ensure adequate level of safety allowing composite action of steel and concrete, but also to control structural behavior along with sufficient ductility. Bond can be achieved between steel bars and concrete in an RC member the mechanical action of ribs of the deformed bar and surrounding concrete. Several experimental studies have been completed to investigate average local and macro response of the bond, and according to these studies several bond models have been successfully reported [1, 2, 3]. Different parameters such as concrete cover, concrete strength, casting direction, elasticity and shape of the rebar were studied while investigating bond behavior [1, 4].

Several experimental studies have been showed the influence of deformation patterns and rib geometry on bond [5, 6, 7]. Local cracks which are formed around deformed bar inside the concrete, affect the bond and those local cracks were successfully observed and schematic diagram represents these cracks was established by (Y. Goto, 1971) [7]. Studies have shown that the aggregate surface area has a significant influence on the mechanical properties of mortar and thereby the bond between deformed bars and surrounding concrete [8].

Based on these global observations, there was a need to simulate bond behavior to predict the performance of the reinforced concrete structures in the future. Meanwhile, computer performance has dramatically improved so that the simulation by modelling the shape of the deformed steel bars becomes possible. It was found that to simulate macroscopic bond behavior; it is necessary to introduce interfacial transition zone around the deformed bar [9]. This porous zone is characterized by reduced tensile strength and elastic modulus of concrete by half due to wall effect [9, 10]..

Recently, using digital image correlation (hereafter referred to DIC) technique to analyze material properties such as concrete has become easier [11, 12, 13]. Study of the bond at the local level with direct observation and measurements of the displacements near the ribs of the deformed bar and thereby strains were recently introduced by using DIC software. Few studies on this technique have shown the applicability of using such a technique in standard pull-out tests in which one deformed bar is embedded centrally in concrete specimen [14].

Regarding this substance, it is very important to investigate the bond behavior in RC experimentally by using advanced techniques, which will enhance our understanding regarding all bond cases and help to improve current bond models in the future. In this study, a powerful experimental procedure for monitoring, and examining of such local deformations between ribs of steel rebar and concrete or cement paste has been proposed, which helps in better understanding of local bond behavior between rebar and concrete, and the mechanisms which govern this behavior. This study also aims to investigate the deformational state under keeping the loading, resulting in that it enables to capture the deformations under the corresponding load levels quantitatively.

2 EXPERIMENTAL PROGRAM

2.1 Materials

Two types of specimens were used in this study; one with concrete and the other with cement paste. Ordinary Portland Cement (OPC) was used as the main binder for both. The water-to-cement ratio (w/c) was 60% by weight. The maximum size of aggregate was 20 mm as generally used in construction projects in Japan. Three standard cylinders of size (100 mm x 200 mm) were used to determine the compressive strength of the concrete and cement paste. Details of the concrete and cement paste mixtures and their strengths are given in Table 1.

Table 1: Mix proportions and fresh properties of concrete and cement paste

	Concrete	Cement Paste
Water (kg/m^3)	180	654
Cement (kg/m^3)	300	1090
Sand (kg/m^3)	807	-
Gravel (kg/m^3)	1008	-
Slump (mm)	90	-
Compressive Strength (MPa)	28.5	24

Table 2: Material properties of reinforcing bar

Diameter (mm)	Yield strength (MPa)	Modulus of Elasticity (MPa)
25	490	190,000

Screw deformed steel bars 25 mm in diameter and 2200 mm long were used for all specimens. Half of the steel bar cross-section was removed to allow directly observation of the rebar ribs and to investigate local deformations between the ribs. This was thought useful for clarifying and understanding behavior in this zone. A metal laser cutting machine was used to remove half of the cross-section (as shown in Figure 1). The properties of the reinforcement are listed in Table 2.

2.2 Test specimen

The set-up of this experiment was performed according to a standard axial tension test and it is shown in Figure 2. Two types of specimens were examined: concrete specimens and cement paste specimens. The test specimen was basically a rectangular of 150 mm x 150 mm x 150 mm. A single screw-shaped reinforcing bar (rebar) was cast in the center of the cross-section. The length of the specimen was determined to be limited as much as possible so that specimen could be easily handled, but at the same time long enough to include a 'window' without affecting the stress distribution. For direct monitoring and observation purpose of the interface between the rebar and the concrete or cement paste, the specimen formwork was not completely casted with concrete or cement paste. Instead, a trapezoidal void was formed during casting by using a 'foam' material (refer to Figure 3). This 'window' might affect the stress distribution inside the specimen due to breaking the symmetry of the specimen. On the other hand, it has an advantage which is allowing close and continuous monitoring and observation of the displacements in the concrete and cement paste. It was found that the height of the 'window' should equal to one-fifth of the total height of the specimen to minimize as much as possible the stress disturbance while at the same time allowing appropriate lightening of the area targeted [9].

2.3 Mixing, casting and curing

The bars were placed horizontally and in the center of the formworks. Concrete or cement paste was poured vertically. Lubricating oil was applied on all sides of the molds for easy removal of specimens. Concrete was poured carefully from the top and segregation was avoided. Needle vibrator was used to compact the concrete. After twenty four hours the specimens were demolded and cured by being covered with wet cloth from just after removing the form to the day before the test in order to avoid the drying shrinkage.

2.4 Technical requirements

Digital image correlation (DIC) system was used to obtain local deformations such as displacements from high-resolution video recorded during experiments (GOM Correlate [15]). Prior to run the DIC analysis, images had to be extracted from the video with same parameters; the software FFmpeg was used for that process [16].

GOM correlate can clearly identify areas of a measurement image which contains sufficient image information for analysis, and then identify the same areas in other images. These areas called "square facets," and the software recognizes them by their stochastic pattern structure, and the facets must have a unique gray level distribution which can be clearly identified. The term stochastic pattern means that it is a pattern that is as random as

possible. This can be achieved by spraying the surface of the concrete or cement paste with a thin layer of white paint then black dots as shown in Figure 4.

High-resolution video camera (4k) was used for recording during the test. The resolution of the camera (i.e., the number of pixels per area in an image) is 3840 pixels × 2160 pixels, to catch very local deformations. The distance between the lens and the surface of the specimen was 10 cm.

2.5 Test procedure

The testing apparatus for axial tension loading is shown in Figure 5. The tensile force was applied to the specimen which was set horizontally on the floor. One end of the deformed bar protruding from the end of the specimen was fixed to the steel frame, which was designed especially for this test, and the tensile force was applied to another end of the bar. The thickness of the steel frame was designed to make the specimen floating and not touching the floor when it is fixed with the steel frame.

As shown in Figure 2, three contact-type displacement transducers (LVDT) were used at different positions (1, 2 and 3) to determine the relative displacement between the deformed bar and the specimen. Actual relative displacement (R.D) between the matrix and the rebar was calculated according to Eq. 1.

$$R.D = \delta 1 - \delta 2 - \delta 3 \quad (1)$$

Where, δ represents the values measured by the LVDTs.

For all tests, the loading rate was kept to about 1 kN/5s by adjusting the speed of the hydraulic pump, and the images were extracted from the video as one image every five seconds to match loading rate.

3 RESULTS AND DISCUSSIONS

Figure 6 shows the load relative displacement response of concrete and cement paste specimens with half cross-section deformed bars. Loading was interrupted twice during the test to verify correct functioning of the system and to check the status of cracks inside the window. The maximum load (Fmax) was about 125 kN because of the reduced cross-section of the deformed bar, which reduced the yield point of the steel to almost half its value which was supposed to be 250 kN.

After finishing the loading process and while monitoring and analyzing the video for the concrete specimen, it was noticed that at the vicinity of the deformed bar, the concrete parts had an intention to move in the direction perpendicular to the rebar which causing opening also between concrete and rebar, not only sliding. So it is important to investigate the opening-sliding relationship between the deformed bar and surrounding matrix. For that purpose, 15 points were chosen, at the tips of the ribs and also between the ribs, to investigate the relative displacements between the deformed bar and surrounding matrix in X-direction and Y-direction (as shown in Figure 7).

3.1 Investigating horizontally along the deformed bar

Figure 8 shows the opening-sliding relationships between concrete and steel bar for 15 points. These points were chosen at distance (1 mm – 1.5 mm) from the surface of the deformed bar. The relationships are divided into four graphs (a, b, c and d). Graph (a) present the points at the tips of the ribs to the left of the central line of loading, graph (b) those at the tips of the ribs to the right of the central line of loading, graph (c) those between the ribs to the left of the central line of loading and graph (d) those between the ribs to the right of the central line of loading.

It is clear that the opening between the concrete and rebar at the tip of the rib (graph (a) and (b)) is higher than those between the ribs (graph (c) and (d)). Also, it is clear that the sliding between the concrete and the rebar at the tips of the ribs (graph (a) and (b)) is higher than those among the ribs (graph (c) and (d)). Here, the main cracks developed and appeared outside the window to the left side of point 1 and to the right side of the point 15. So

it can be noticed that at the same sliding, the opening at points which are far from the main crack is higher than that near the main crack. Also, at points far from the main crack, the opening is much higher than the sliding comparing to points near the main crack. In addition, it can be concluded that, far from the main crack, the inclination or the slope of the opening-to-sliding ratio changes by (2:1), meanwhile, near the main crack it shifts from (2:1) to (1:3).

Figure 9 shows also the opening-sliding relationships for 15 similar points but for the cement paste specimen. The categorization is same as for the concrete case (a, b, c and d). In cement paste case, the trend is much easier to catch than the concrete. It is very clear that the relation between the opening and the sliding is roughly linear; and the slope of their ratio changes from 3:1 far from the main crack to 1:1 near the main crack. It can be seen that the opening between the cement paste and the rebar at the tip of a rib is higher than between the ribs. Also, at the same sliding, the opening at points which are far from the main crack is higher than that near the main crack. Therefore, the difference between the concrete and the cement paste; in case of cement paste, the difference between points is somehow constant because the softening of cement paste is very low comparing to the concrete. As a result, few cracks can be found in cement paste rather than concrete but these cracks are very much wider than those in concrete. So the opening and sliding of the cement paste depends on the displacement of the crack, because it is a homogenous material. Meanwhile, in the case of concrete, opening and sliding depends on the number of cracks and the stiffness of the coarse aggregate particles as it is non-homogenous material.

3.2 Investigating vertically along the deformed bar

To investigate deeply the surrounding zone around the deformed bar, two points were chosen to be more investigated for the same opening-to-sliding relationship. These points were chosen to be far as much as possible from the main cracks to not be affected by the characteristics of these cracks. These two points were investigated for the relative displacements in X and Y directions between the rebar (fixed point) and the surrounding concrete or cement paste at 1 mm, 3 mm, 5 mm, 7 mm and 9 mm from the face of the rebar (refer to Figure 10). For the concrete case, points 3 and 4 were chosen for the investigation and for the cement paste case, points 2 and 3 were chosen. Figure 11 shows two graphs for the opening-sliding relationships of points 3 and 4 in the concrete case, while Figure 12 shows the same relationships but for points 2 and 3 in the cement paste case. From Figure 11, it is clear that both points 3 and 4 at all locations inside the concrete are moving almost like a rigid body till 70% of F_{max} (at 0.25 mm sliding), while after 70% of F_{max}, the trend of the each point at 1 mm is totally different from other locations. This deformational behavior could be explained by the shear interface zone deterioration which is concentrated around 1-2 mm from the face of the rebar.

It is also clear from both graphs of the concrete (as shown in Figure 11), as going far from the surface of the rebar, the opening between concrete and the rebar increases more than sliding, which means that with going far vertically from the rebar, the opening becomes more dominant. This could be explained by Figure 12, as the internal stress inside concrete progresses, there are numerous micro cracks which occur between these locations (1 mm - 3 mm - 5 mm - 7 mm - 9 mm) due to the heterogeneity of the concrete material, so the opening between the rebar and concrete at location of 5 mm will be a cumulative opening from 0 mm till 5 mm.

4 CONCLUSIONS

This paper presents an experimental study of axial tension tests of deformed bars in concrete and cement paste specimens. For direct monitoring and observation purpose of the interface between the rebar and the concrete or cement paste, each specimen have

a trapezoidal which was formed during casting. DIC analysis was used to visualize local behavior within the concrete or cement paste. The following conclusions have been drawn from this work:

1. The DIC technique was used to find the normal stresses at the interface between concrete and steel that result in opening between concrete or cement paste and rebar.
2. The opening between concrete and rebar at the tip of a rib is greater than between ribs. Also, it was found that sliding between concrete and rebar at the tip of a rib is higher than between ribs. For a certain amount of sliding, the opening at points far from the main crack is greater than at those near the main crack. At points far from the main crack, the opening is much higher compared to the sliding, in contrast with points near the main crack.
3. In the case of cement paste, the relation between opening and sliding is roughly linear and stable. Moreover, the opening between cement paste and rebar at the tip of a rib is higher than between the ribs.
4. As going far from the surface of the rebar, the opening between concrete and the rebar increases more than sliding, which means that with going far vertically from the rebar, the opening becomes more dominant.

REFERENCES

[1] H. Shima, L. L. Chou, and H. Okamura, "Micro and macro models for bond in reinforced concrete," Journal of the Faculty of Engineering, 39(2), 133-194, 1987.

[2] R. Eligehausen, E. P. Popov, and V. V. Bertero, "Local bond stress-slip relationships of deformed bars under generalized excitations," University of California, Report no. UCB/EERC-83/23 of the National Science Foundation, 1983.

[3] L.A. Lutz and P. Gergely, "Mechanics of bond and slip of deformed bars in concrete," ACI Journal, 64(11), 711-721, 1967.

[4] B. S. Hamad, "Bond strength improvement of reinforcing bars with specially designed rib geometries," ACI structural journal, 92(1), 3-13, 1995.

[5] B. B. Broms, "Technique for investigation of internal cracks in reinforced concrete members," ACI Journal, Proceedings, 62(l), 35-44, 1965.

[6] D. Watstein, and R. G. Mathey, "Width of Cracks in Concrete at the Surface of Reinforcing Steel Evaluated by Means of Tensile Bond Specimens," ACI Journal, 56(1), 47-56, 1959.

[7] Y. Goto, "Cracks formed in concrete around deformed tension bars," ACI Journal, 68(4), 244-251, 1971.

[8] C. F. Goble, and M. D. Cohen, "Influence of aggregate surface area on mechanical properties of mortar," ACI materials Journal, 96(6), 657-662, 1999.

[9] H.M. Salem, and K. Maekawa, "Pre-and post-yield finite element method simulation of bond of ribbed reinforcing bars," Journal of Structural Engineering, 130(4), 671-680, 2004.

[10] L. Eddy, K. Matsumoto, K. Nagai, and Z. Wang, "Investigation of effect of local bond model on crack propagation in RC using 3D RBSM," Proceedings of Euro-C 2018, Bad Hofgastein, Austria, 26 February - 1 March 2018.

[11] D. Corr, M. Accardi, L. Graham-Brady, and S. Shah, "Digital image correlation analysis of interfacial de-bonding properties and fracture behavior in concrete," Engineering Fracture Mechanics, 74(1), 109-121, 2007.

[12] J. Kozicki, and J. Tejchman, "Experimental investigations of strain localization in concrete using Digital Image Correlation (DIC) technique," Archives of Hydro-Engineering and Environmental Mechanics, 54(1), 3-24, 2007.

[13] S.G. Shah, and J.M Chandra Kishen, "Determination of fracture parameters of concrete interfaces using DIC," Fracture Mechanics of Concrete and Concrete

Structures, 1208-1215, 2010.
[14] O. Leibovich, A. N. Dancygier, and D. Z. Yankelevsky, "An Innovative Experimental Procedure to Study Local Rebar-Concrete Bond by Direct Observations and Measurements," Experimental Mechanics, 56(5), 673-682, 2016.
[15] GOM Correlate (2016), "Inspection and Testing Guide," available at http://www.gom.com/index.html. Accessed Aug. 2014.
[16] F. Bellard, M. Niedermayer, FFmpeg, http://ffmpeg.org, 2012

Figure 1: Cutting half of the cross section of the rebar by laser cutting

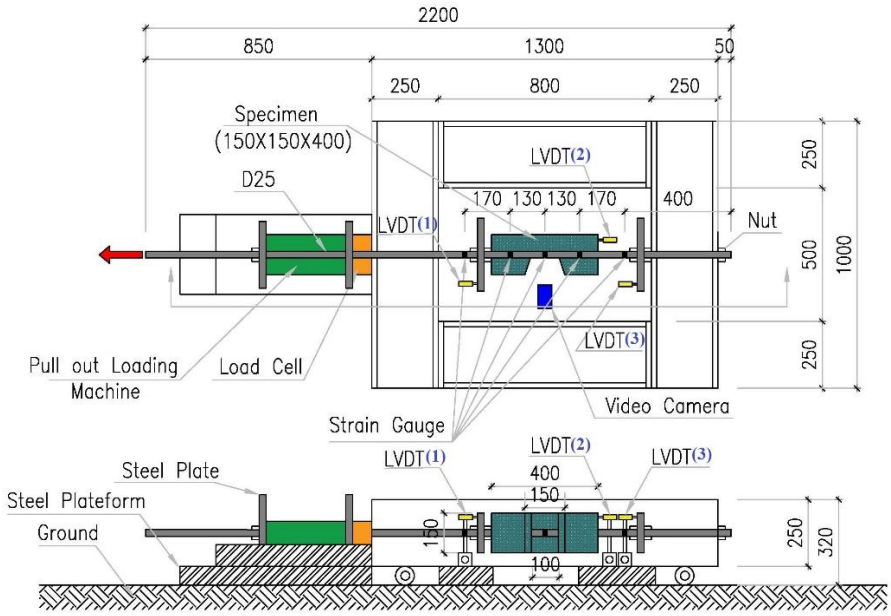


Figure 2: The set-up of the experiment

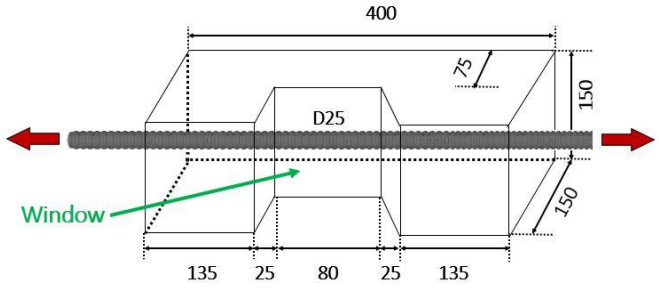


Figure 3: Lay-out of the test specimen (units: mm)

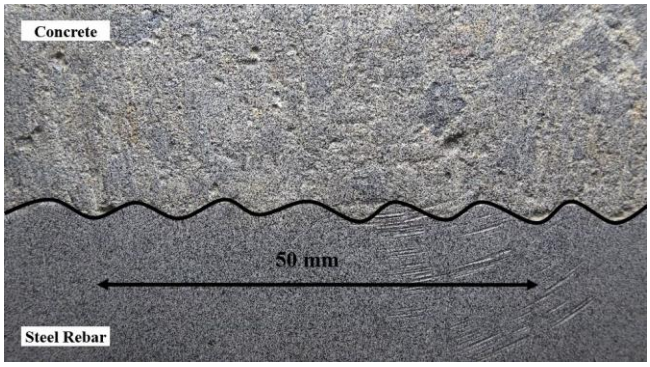


Figure 4: Sprayed specimen of the photographed area of interest

Figure 5: Testing apparatus and steel frame of the test

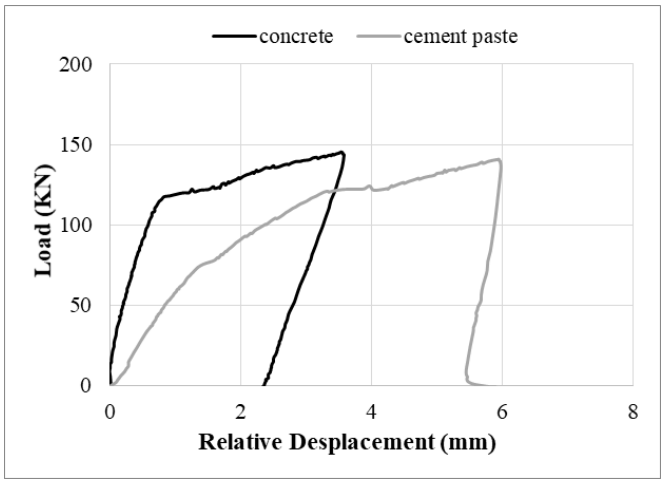
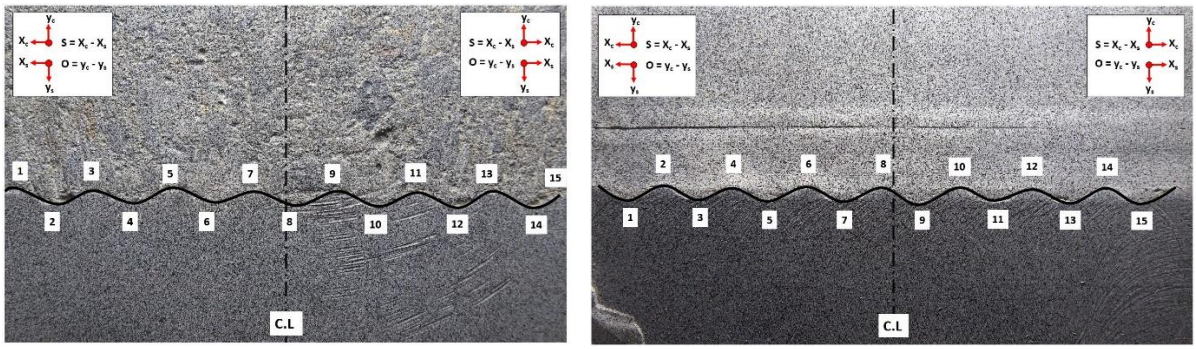


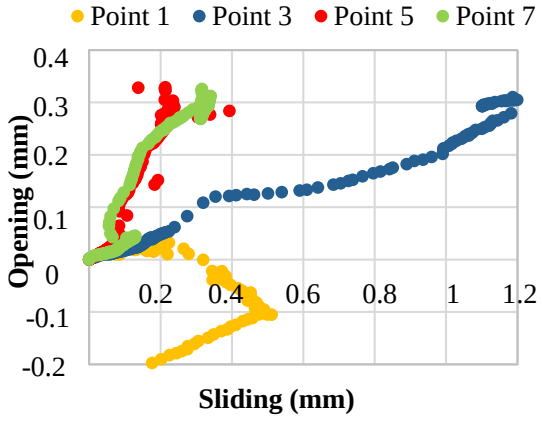
Figure 6: Load-relative displacement relationships

Concrete **Cement Paste**

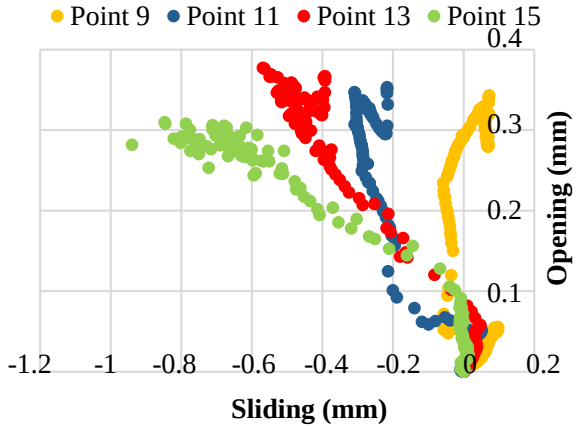


Note: S is sliding between matrix and rebar, O is opening between matrix and rebar, X_c, Yc are displacements of matrix in X, Y directions respectively, X_s,Y_s are displacements of rebar in X,Y directions respectively

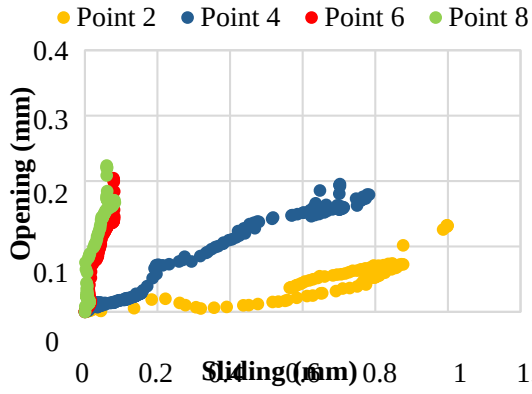
Figure 7: Selected points at the tips of ribs and between ribs for concrete and cement paste



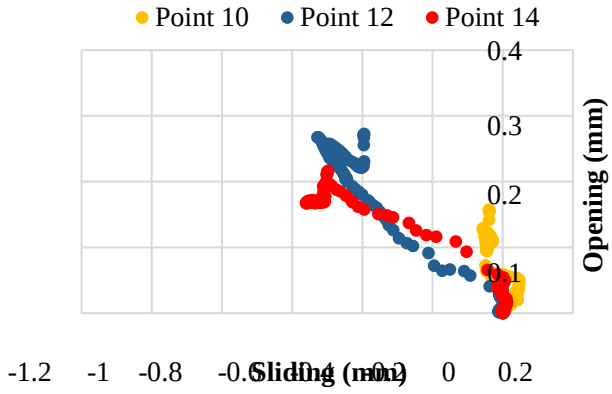
(a) Points at tips of ribs to the left of central line



(b) Points at tips of ribs to the right of central line

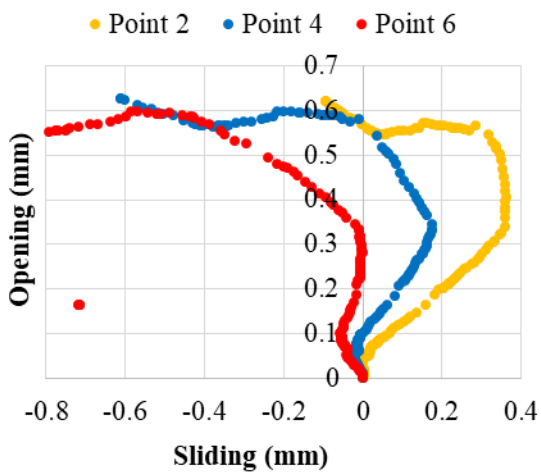


(c) Points between ribs to the left of central line

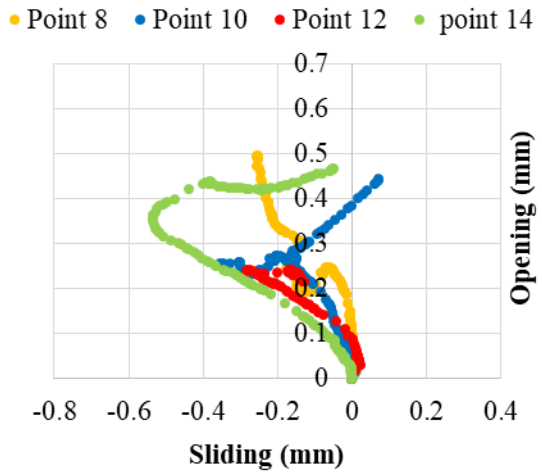


(d) Points between ribs to the right of central line

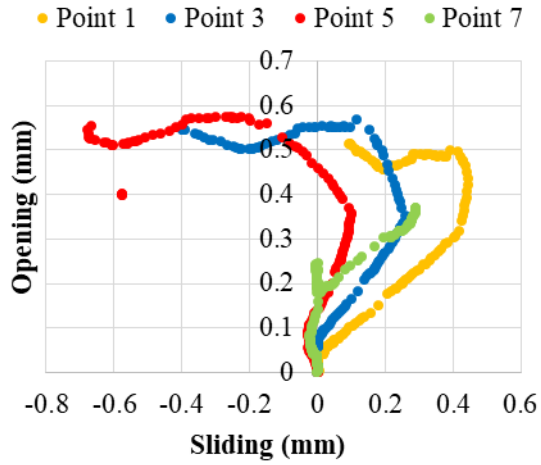
Figure 8: Opening-sliding relationships in concrete



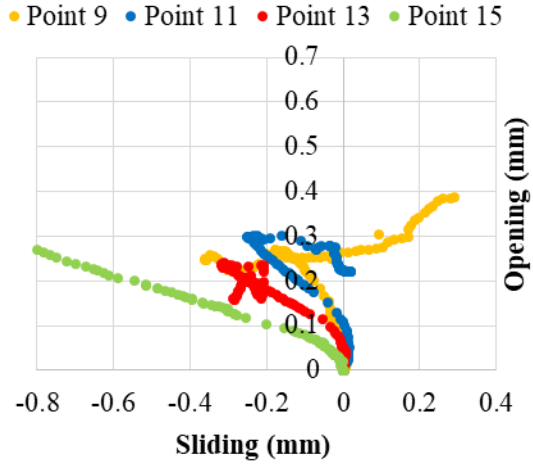
(a) Points at tips of ribs to the left of central line



(b) Points at tips of ribs to the right of central line



(c) Points between ribs to the left of central line



(d) Points between ribs to the right of central line

Figure 9: Opening-sliding relationships in cement paste

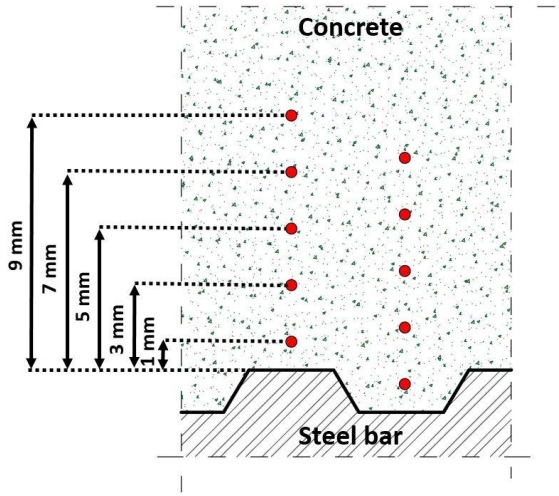


Figure 10: Lay-out for two points with different positions (1-3-5-7-9 mm).

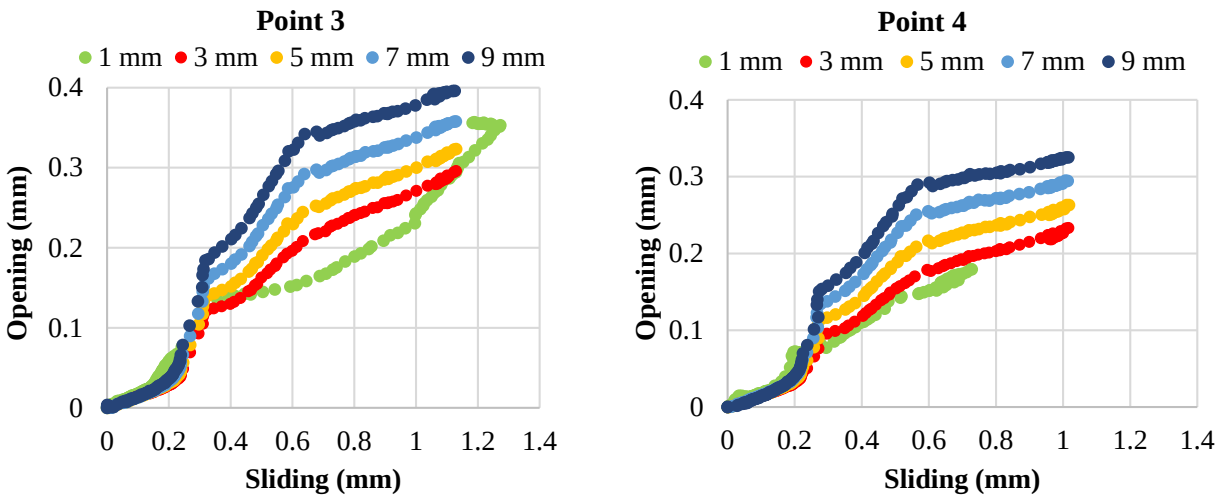


Figure 11: Opening-sliding relationship in concrete for two points at different positions (1-3-5-7-9 mm).

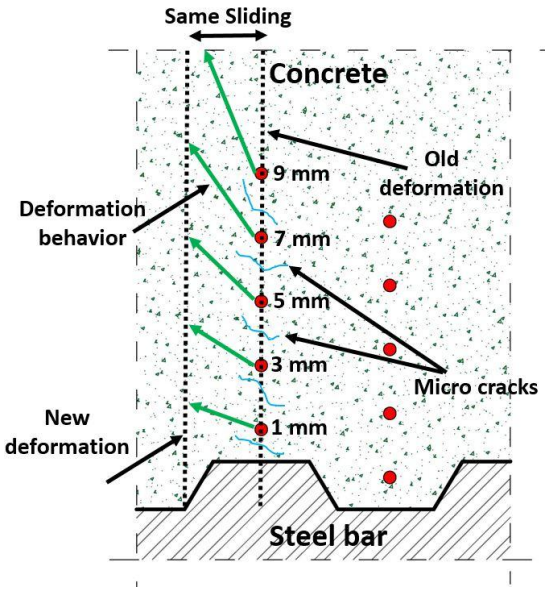


Figure 12: Schematic of opening-sliding behavior in concrete with going vertically from the rebar.

Received: 7 January 2019 | Revised: 16 May 2019 | Accepted: 5 July 2019

DOI: 10.1002/fam.2751

RESEARCH ARTICLE

WILEY

Investigation on flammability of rigid polyurethane foam-mineral fillers composite

Anuja Agrawal[1] | Raminder Kaur[1] | Ravinderjit Singh Walia[2]

[1] Department of Applied Chemistry and Polymer Technology, Delhi Technological University, Delhi 110042, India

[2] Department of Mechanical Engineering, Delhi Technological University, Delhi 110042, India

Correspondence
Raminder Kaur, Department of Applied Chemistry and Polymer Technology, Delhi Technological University, Delhi 110042, India.
Email: rkwaliadtu@gmail.com

Funding information
Defence Research and Development Organisation, Grant/Award Number: CC/TM/ERIPR/GIA/15-16/029

Summary

This study investigates the incorporation of castor oil–based rigid polyurethane foam with mineral fillers feldspar or kaolinite clay in order to enhance the mechanical, thermal, and flame retardant properties. Influence of mineral fillers on the mechanical strength was characterized by compressive strength and flexural strength measurement. Thermogravimetric analysis (TGA) was performed to diagnose the changes in thermal properties, while cone calorimeter test was performed to ascertain the flame retardancy of the mineral filler–incorporated rigid polyurethane foam composites. Results showed that the foams incorporated with mineral filler demonstrated up to 182% increase in compressive strength and 351% increase in flexural strength. Thermal stability of these composite foams was also found to be enhanced on the incorporation of kaolinite clay filler with an increase in 5% weight loss temperature ($T_{5\%}$) from 192°C to 260°C. Furthermore, peak heat release rate (PHRR), total heat release (THR), smoke production rate (SPR), and total smoke release (TSR) were also found to decreased on the incorporation of mineral filler in the rigid polyurethane foam. So mineral fillers are ascertained as a potential filler to enhance the mechanical, thermal, and flame retardant behaviors of bio-based rigid polyurethane foam composites.

KEYWORDS

degradation temperature, heat release rate, mechanical strength, smoke production, thermal stability

1 | INTRODUCTION

Rigid polyurethane foam (RPUF) has its application in a wide variety of fields such as construction, transport, thermal insulation, packaging of electronic goods, and radomes.[1,2] High-density urethane cellular foam shows little deflection on applying heavy structural loads during construction applications. It also has applications in hand-carved models, prototypes, Computer Numerical controlled (CNC) -machined topographical maps and industrial patterns.[3] Polyurethane (PU) foams derived from renewable sources are requirement of modern world due to various issues related to petroleum-based foams.[4,5] PU foams are prepared by using polyols and polyisocyanate as raw materials. Although polyols have been successfully derived from renewable sources,[6,7] studies reveal that the functional properties (compressive strength, thermal stability, and flammability) of these renewable source–based foams are inferior to the foam derived from petroleum-based raw materials. Thus, numerous studies have been conducted to improve these functional properties of renewable sources–based foams by incorporating different fillers.[8] Additives such as fibers or particulate fillers are used as reinforcement to increase the compressive strength, sound attenuation, thermal stability, and flame retardancy and to decrease density and water absorptivity of the PU foam. Serban et al[9] investigated the influence of the glass fiber incorporation on the mechanical properties of PU foam in terms of the variations in fiber length (12.5-50 mm) and fiber mass content (5%-20%). Results concluded that the glass fiber reinforcement increases the

 © 2019 John Wiley & Sons, Ltd.

elastic modulus more than two folds as compared with unreinforced PU foam on the incorporation of 20% glass fiber with fiber length of 25 mm. Xue et al[10] developed a lignin-based rigid polyurethane foams (LRPF) reinforced with different weight ratios (1-5 wt%) of pulp fiber. Addition of pulp fiber in the LRPF resulted in greater cell size, lower density, and better thermal stability of the foam.

Apart from these fibrous fillers, some particulate fillers such as finely divided mineral fillers have additional advantages of providing high heat-deflection temperature, low thermal expansion, good moldability, and low cost, which are the requirement of good RPUF. These additional advantages are strongly affected by the filler-to-resin ratio, extent of dispersion, and particle size of the filler.[11,12] Different types of mineral fillers such as calcium carbonate,[13] talc,[14] and nanoclay[15,16] have been considered to improve the properties of PU foams. Czupryński et al[17] incorporated various particulate fillers in rigid polyurethane-polyisocyanurate foams and analyzed their effect on the properties of the rigid foam. The addition of aluminum hydroxide and talc into the foam improved its various functional properties such as increase in compressive strength and closed cell content and decrease in the apparent density, flammability, and brittleness, while the addition of chalk, borax, and starch reportedly deteriorated its properties. Nik Pauzi et al[18] considered modified diaminopropane montmorillonite (DAP-MMT) nanoclay (0-10 wt%) as a promising reinforcement material to incorporate in the palm oil-based PU foam. An increase in compressive strength and thermal stability was observed on increasing the concentration of DAP-MMT up to 4%. Zheng et al[19] introduced organically modified montmorillonite (OMMT) nanoclay as a filler in the polyurethane matrix to enhance the flame retardancy of RPUF. Phosphorus-based flame retardant such as ammonium polyphosphate (APP) and triphenyl phosphate (TPP) were also incorporated to enhance the char-forming property of RPUF. It was observed that foam with 8% APP, 4% TPP, and 5% OMMT exhibited maximum increase in combustion duration time and a decrease in heat release rate (HRR) and total smoke production. Zieleniewska et al[20] introduced the egg shell waste (0-25%) into RPUF and investigated its effect on the material properties of the foam. Results indicated an increase in the compressive strength and density, better thermal stability, low water absorption, and high dimensional stability in wet conditions. Zhu et al[21] incorporated cellulose microfibers and nanoclay in the RPUF to alter the cellular structure of the foams. The compressive strength of foam was found to be increased by using these fillers as reinforcement. Atabek[22] analyzed the effect of calcium hypophosphite on the properties of some thermoplastic polymers including polyurethane. It was observed that the thermal stability and flame retardancy of the polyurethane was increased on the incorporation of the filler. The synergistic effect of some mixed inorganic flame retardant has also been studied. Mixture of expandable graphite or red phosphorus with guanidinium phosphate was employed as a flame retardant, and a noticeable improvement in the flame retardancy of the RPUF was observed.[23] Additionally, some flame retardants such as aluminum diethyl phosphinate,[24] hydroxyl-terminated monomer containing phosphorus and nitrogen, tri(N, N-bis-(2-hydroxy-ethyl) acyloxoethyl) phosphate (TNAP),[25] or modified urea-formaldehyde resin[26] has also been used to improve the flame retardancy of the polyurethane. Summary of different fillers used in PU matrix for enhancing various properties is presented in Table 1.

Although numerous studies have already been conducted for the properties enhancement of RPUF using mineral fillers such as calcium carbonate, talc, and nanoclay, some other low-cost natural mineral fillers such as feldspar and kaolinite clay have not been explored so far to enhance the functional properties of bio-based RPUF. Feldspar is an alumina-silicate mineral and being used as fillers in the adhesive, rubber, and paint industries to improve the strength, toughness, and durability of the final products. Beneficial properties of feldspar include high chemical inertness, high resistance to abrasion, and low viscosity at high filler loading.[27] Kaolinite clay is used by industries as a filler to reduce the cost and to enhance the strength. Kaolinite clay is chemically inert, nonabrasive with low plasticity and apparent bulk density as well as high refractory properties.[28,29] So, in this study, mineral fillers kaolinite clay and feldspar have been used as reinforcement and their effect on the mechanical, thermal, and flame retardant behaviors of bio-based RPUF has been investigated.

2 | EXPERIMENTAL

2.1 | Materials

RPUFs were prepared by using Castor oil (99.9%), diphenylmethane diisocyanate (PMDI, 30-32 wt% NCO), *n*-Pentane (99.5%), silicon oil, glycerol (99%), and Catalyst Dabco 33-LV (1,4-Diazabicyclo[2.2.2] octane solution, Sigma-Aldrich). Kaolinite clay (particle size 100-500 nm with above 98% kaolin contents) and feldspar (particle size nanomicron to submicron range and maximum particle size 3-4 μm) were purchased from Zirox technologies, Bewar, Rajasthan. All the reagents were of analytical grade and used as supplied.

2.2 | Methods

2.2.1 | Foam preparation

The RPUFs were obtained by the same method as reported in the previous literature.[30,31] In brief, the procedure was as follows:

The modification of castor oil was performed under the inert atmosphere of nitrogen using 2:1 ratio of castor oil to glycerol at the temperature of 180°C to 200°C for 4 to 5 hours. The predetermined quantity of the filler was added to the modified polyol of hydroxyl value 400 ± 30 mg KOH/g taken in a paper cup. Prior to addition, the fillers were surface treated with 1% solution of 3-aminopropyl triethoxysilane (APTES) in acetone/water 50/50 by volume using the method described in literature,[32] to enhance the adhesion between matrix and filler. Then calculated amount of other ingredients such as catalyst (Dabco 33-LV), surfactant (Silicon Oil), and blowing agent (*n*-Pentane) was added to the contents of the paper cup and thoroughly mixed at 1500 rpm by mechanical stirrer for 2 minutes under controlled temperature conditions at 35 ± 2°C to form polyol premix.

TABLE 1 Fillers used in PU matrix for enhancing various properties

S. No.	Matrix Material	Properties to Enhance	Filler Incorporated	Reference
1.	PU foam	Mechanical properties	Glass fiber	Serban et al[9]
2.	Lignin-based RPUF	Mechanical and thermal properties	Pulp fiber	Xue et al[10]
3.	Polyurethane	Thermal properties	Calcium carbonate	Gao et al[13]
4.	Rigid polyurethane foam	Mechanical properties	Talc powder	Alavi Nikje et al[14]
5.	Rigid polyurethane foam	Mechanical properties	Microsphere and nanoclay	Fan et al[15]
6.	Rigid polyurethane foam	Thermal conductivity	Nanoclay	Estravís et al[16]
7.	Polyurethane-polyisocyanurate foams	Mechanical properties, flammability	Aluminum hydroxide, talc, borax, starch, chalk	Czupryński et al[17]
8.	Palm oil-based PU foam	Mechanical properties, thermal stability	diaminopropane montmorillonite (DAP-MMT) nanoclay	Nik Pauzi et al[18]
9.	Rigid polyurethane foam	Flame retardancy	Organically modified montmorillonite, ammonium polyphosphate (APP) and triphenyl phosphate (TPP)	Zheng et al[19]
10.	Rigid polyurethane foam	Mechanical properties, thermal stability, water absorption	Egg shell waste	Zieleniewska et al[20]
11.	Rigid polyurethane foam	Mechanical properties	Cellulose microfibers and nanoclay	Zhu et al[21]
12.	Thermoplastic polyurethane	Thermal stability and flame retardancy	Calcium hypophosphite	Atabek[22]
13.	Rigid polyurethane foam	Flame retardancy	Mixture of expandable graphite or red phosphorus with guanidinium phosphate	Chen et al[23]
14.	Polyurethane	Flame retardancy	Aluminum diethyl phosphinate	Wang et al[24]
15.	Polyurethane	Flame retardancy	Tri(N, N-bis-(2-hydroxy-ethyl) acyloxoethyl) phosphate (TNAP)	Wang et al[25]
16.	Polyurethane	Flame retardancy	Modified urea-formaldehyde resin	Zhu and Xu[26]

Abbreviations: RPUF, rigid polyurethane foam.

A calculated amount of MDI was then added to the paper cup and mixed at 2000 rpm for 10 seconds. Then, the prepared reaction mixture was poured into a metal mold (100 mm × 100 mm × 25 mm) coated with a releasing agent (silicon oil), and a free rise foam was prepared. To ensure complete curing, the molds were left at room temperature for 72 hours. Different concentrations of the mineral fillers were used to analyze the influence of mineral filler on the properties of RPUFs. RPUF prepared without adding fillers has been considered as reference sample. Table 2 shows the formulation of rigid castor oil-based polyurethane foams modified with different concentration of mineral filler.

TABLE 2 Formulation of RPUF modified with different mineral fillers

Material	Amount (pbw)
Modified polyol	100
Dabco 33-LV	3
n-Pentane (blowing agent)	15
Silicone oil (surfactant)	3
Filler	0-10
MDI	110

Abbreviations: pbw, parts by weight; RPUF, rigid polyurethane foam.

2.3 | Characterization

The cell structure of the samples was characterized with a Hitachi S3700 SEM (scanning electron microscope). Samples were gold sputter coated to render them electrically conductive and then scanned at an accelerating voltage of 15 kV. Image processing was conducted by using in-built software SEM data manager of Hitachi PC-SEMs.

The mechanical properties of the castor oil based rigid polyurethane samples were determined according to the standard procedures. At least three specimens of each concentration were tested and the average value was reported. Compressive and flexural properties of the foam specimen were measured at room temperature using Instron (model No. 3369) universal testing machine (UTM). The compression tests were performed according to the ASTM D-1621. Specimens of dimensions 25 mm × 25 mm × 25 mm were cut from the foam in the in-plain direction and tested for 10% compression. The compressive strength was measured along parallel foam rise direction. The flexural tests were performed according to the ASTM D-790.

Specimens of dimensions 80 mm × 10 mm × 4 mm were cut from foam perpendicular to foam rise direction and loaded under three-point bending along parallel foam rise direction. The rate of crosshead movement was fixed at 5 mm/min for each sample.

Thermogravimetric analysis (TGA) was performed on foam samples utilizing a thermogravimetric analyzer (Perkin Elmer 4000) in Nitrogen atmosphere. The heating rate was 10°C/min with temperature ranging from 30°C to 700°C. All the samples were run in triplicate and the average value is reported.

The response to fire of RPUF samples was examined with a Cone Calorimeter (Jupiter Electronics, India) according to ISO 5660-1 standard at an incident heat flux of 35 kW/m^2. The size of the samples was taken as 100 mm × 100 mm × 20 mm. This instrument is capable of recording the time to ignition (TTI), total heat release (THR), HRR, smoke production rate (SPR), and total smoke production (TSP), etc. All the samples were tested in duplicate, and the average is reported.

3 | RESULTS AND DISCUSSION

3.1 | Mechanical properties

Mechanical properties of PU foam are influenced by several factors such as size and shape of cells along with density of the foam. The foam structure also depends on the particle size and the dispersion of the filler in the reacting mixture. Moreover, the interaction between the filler and polymer matrix is also an important parameter governing the mechanical properties of reinforced PU foam.[17] Figures 1–5 show the change in the mechanical strength and specific mechanical strength (strength/density) of RPUF incorporated with different mineral fillers. Incorporation of these fillers in the RPUF increases the compressive strength of the foam initially, and then a decrease of the property was observed with further increase of filler concentration. The maximum compressive strength was shown on the addition of feldspar at 8% and kaolinite clay at 6% filler concentration, while foams incorporated with 8% feldspar showed the highest compressive strength of 4.98 ± 0.25 MPa, which is 182% greater than that of neat foam (Figure 1). Flexural strength also demonstrated the similar trends, and foams with 6% kaolinite clay showed the highest flexural strength of 3.56 ± 0.18 MPa, which is 351% greater than that of neat foam (Figure 2).

Increasing the filler concentration beyond this limit led to the agglomeration of the filler, which produces only filler-filler interaction, consequently deteriorating the mechanical properties of the foam. It is proved by the distorted structure of the foams at higher filler concentration as shown by SEM images. Similar findings are also reported in the literature.[18] The density of rigid foams also increases on increasing the concentration of mineral fillers (Figure 3). As the density has a remarkable effect on the mechanical strength of the foam, so specific mechanical strength provides the idea about the change in strength due to incorporation of mineral fillers in the RPUF (Figures 4 and 5). Decrease in cell size and subsequent increase in cell number contribute to a superior mechanical strength because the load applied is encountered by more cell walls per unit area of the foam. Furthermore, incorporation of the fillers in the cell walls strengthened the foam structure, consequently increasing the mechanical strength.[33] Furthermore, feldspar possesses more percentage of silica in its composition, which forms the restraining network and enhances the mechanical strength more in comparison with kaolinite clay.[29]

3.2 | Scanning electron microscopy (SEM)

Physical and mechanical properties of RPUF are remarkably affected by the morphology of the foams. Figures 6 and 7 show the SEM images of the fractured surface of the RPUF composites and reveals their cell structure. Neat PU foams showed cell size in the range of 400 ± 20 μm, while incorporation of mineral fillers decreased the cell size to 350 ± 20 μm and 280 ± 20 μm for the foams reinforced with 8% feldspar and 6% kaolinite clay, respectively. This behavior is attributed to the fact that fillers provide more nucleation sites for bubble formation. In addition, mineral fillers do not float easily with the matrix

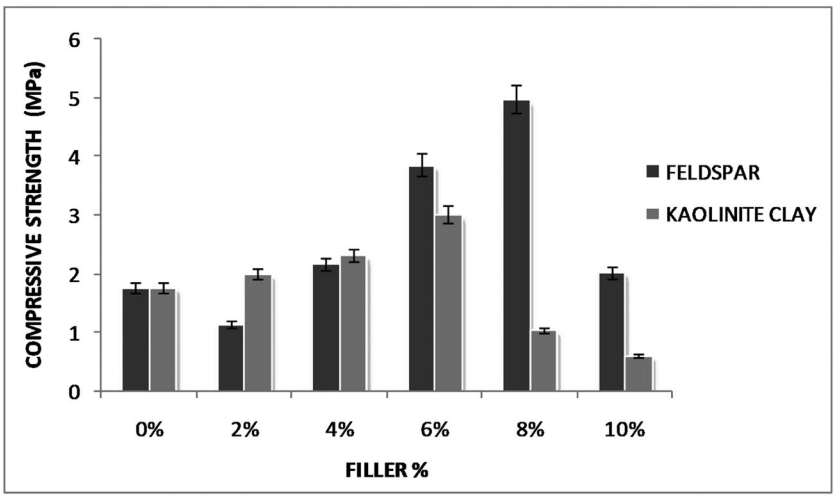


FIGURE 1 Variation in compressive strength with increase in concentration of fillers

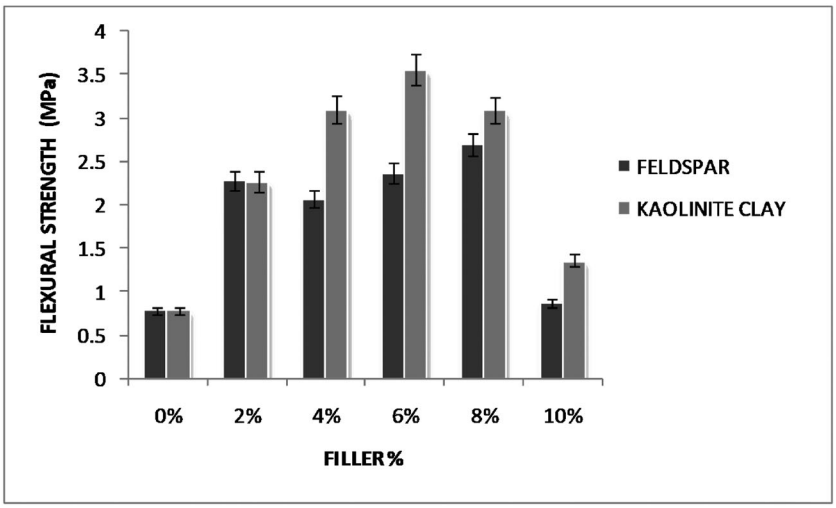


FIGURE 2 Variation in flexural strength with increase in concentration of fillers

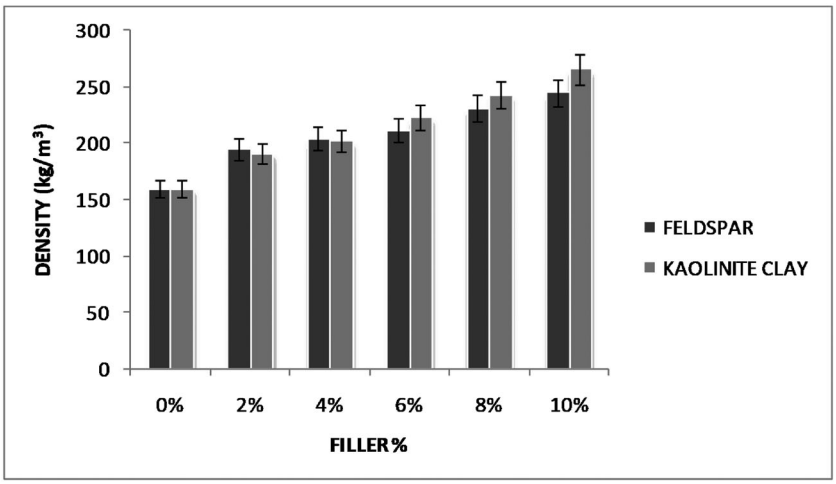


FIGURE 3 Variation in density with increase in concentration of fillers

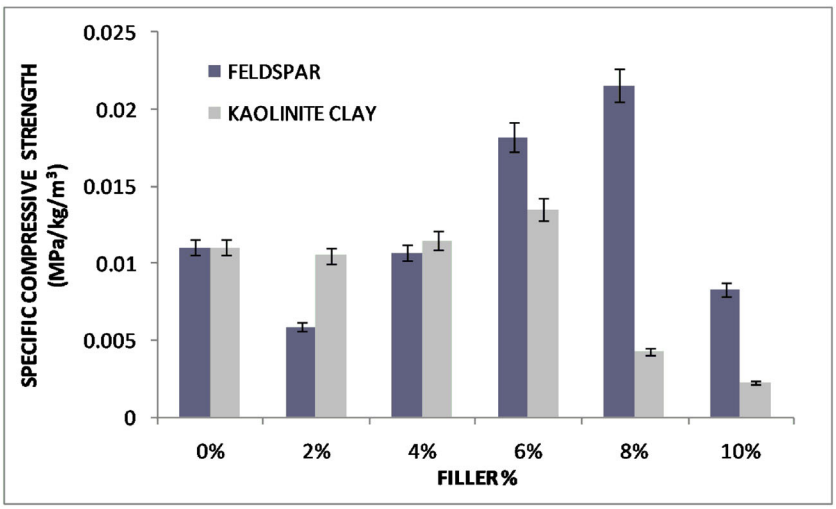


FIGURE 4 Variation in specific compressive strength with increase in concentration of fillers

consequently increasing the viscosity of the matrix while foaming, which causes the resistance to the increase in cell size.

Furthermore, incorporation of filler increases the viscosity of reaction mixture, which prevents the bubble from being coalesced and consequently leading to the decrease in cell size. The reduced cell size comes out with more packed foam structure that results in increased foam density and mechanical strength, which is accredited to the availability of more struts per unit area of RPUF.[18] Increasing the

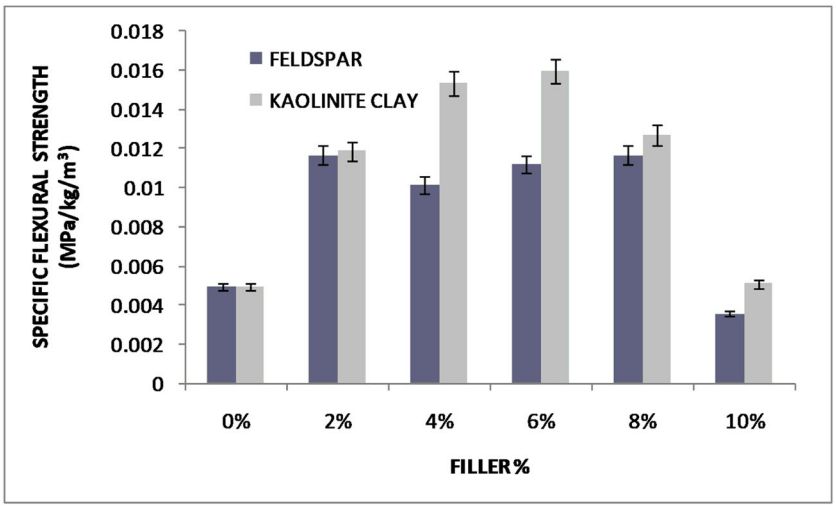


FIGURE 5 Variation in specific flexural strength with increase in concentration of fillers

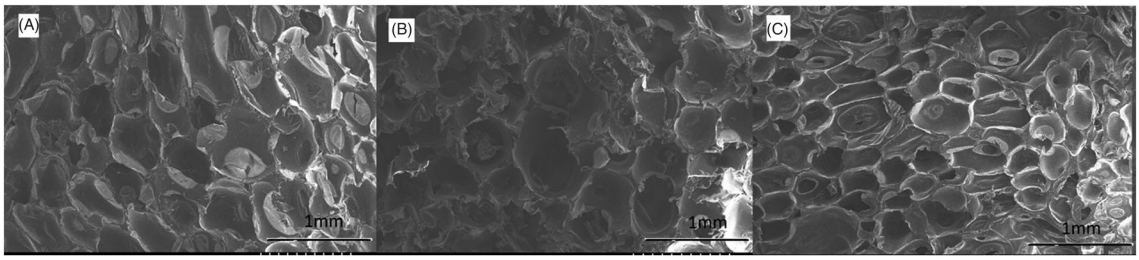


FIGURE 6 SEM micrographs of foam reinforced with A, 0% filler, B, 8% feldspar, and C, 6% kaolinite clay

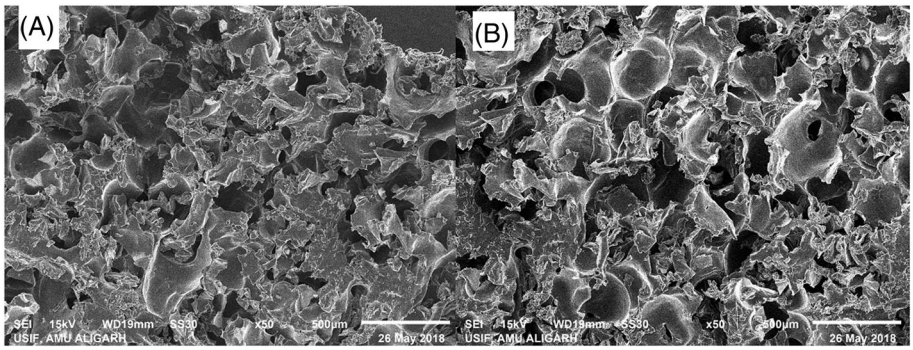


FIGURE 7 SEM micrographs of foam reinforced with A, 10% kaolinite clay and B, 10% feldspar

concentration of fillers beyond 8% for feldspar and 6% for kaolinite clay deteriorates the cell structure that is associated with the agglomeration of mineral fillers after a certain concentration due to the high viscosity of the matrix-filler premix.

3.3 | Thermogravimetric analysis

To evaluate the thermal stability of the prepared rigid foam composites, TGA is conducted under the flow of nitrogen. The TGA and Derivative thermogravimetry (DTG) analysis of mineral filler–incorporated PUF are illustrated in Figure 8A-D, and the parameters are summarized in Table 3.

RPUFs showed two stages of weight loss at 300°C to 350°C and 450°C to 500°C. The first peak at around 300°C to 350°C is attributed to the degradation of the urethane and urea linkages in the hard segments and the decomposition of the flexible phase containing polyol. The second peak observed at 450°C to 500°C is related to the thermolysis of the organic residues and the degradation of soft segments.[34]

Conventionally, the thermal stability of RPUFs is described by the temperatures of 5% weight loss ($T_{5\%}$) considered as the temperature for the onset of degradation.[35] Results showed that 5% weight loss temperature ($T_{5\%}$) increased from 192 ± 2°C to 218 ± 2°C for the foam with 8% feldspar and 260 ± 2°C for the foam with 6% kaolinite clay, then again decreased due to heterogeneity generated by agglomeration of

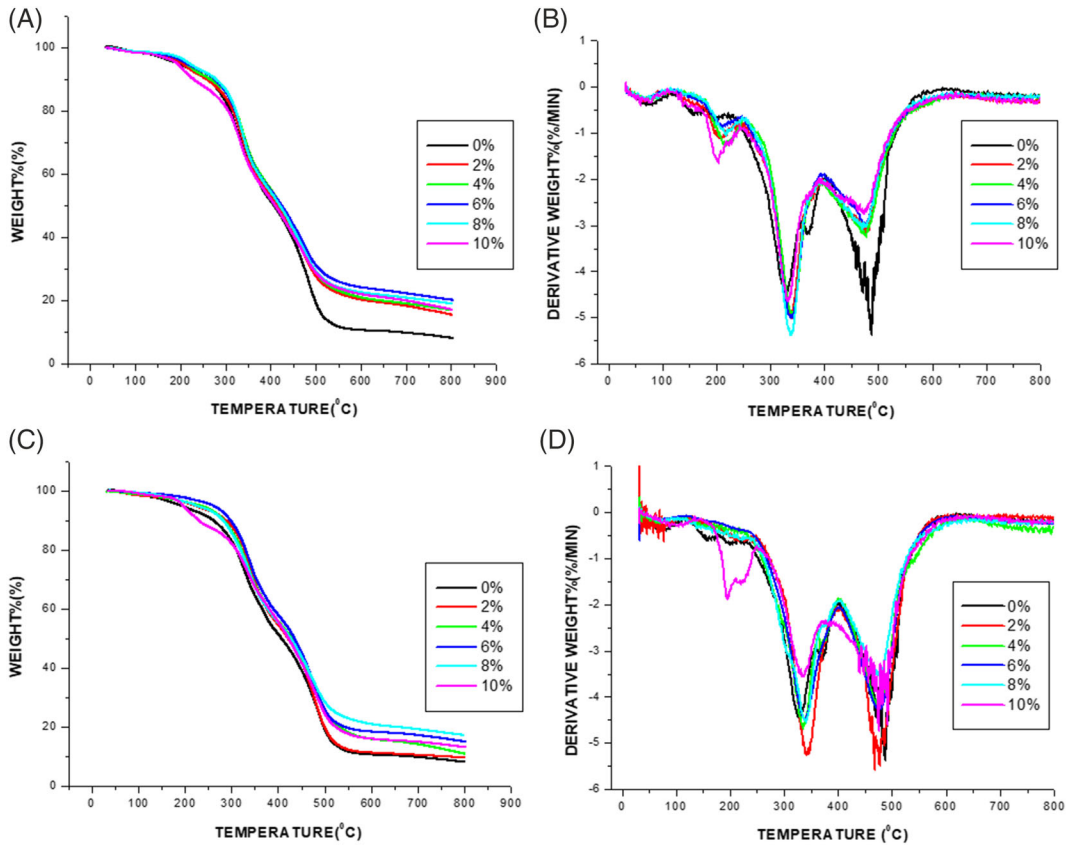


FIGURE 8 A, Thermogravimetric analysis (TGA) curve of rigid polyurethane foam (RPUF) incorporated with feldspar powder, B, DTG curve of RPUF reinforced with feldspar, C, TGA curve of RPUF incorporated with kaolinite clay, and D, DTG curve of RPUF reinforced with kaolinite clay

the filler at higher concentration. T_{max1} increases slightly on the incorporation of mineral filler as mineral fillers execute the function of the thermal barrier that prevents the rapid advancement of the heat and restricts the further degradation. Furthermore, bulk density of the foam material is increased by the exfoliation or the intercalation of kaolinite clay within the RPUF, which restricts the diffusion of volatile decomposition products within the RPUF consequently increasing the thermal stability of the foam. Owing to endothermic decomposition of the mineral fillers, they absorb heat and therefore keep the enclosing polymer cooler. Large aspect ratio of the kaolinite clay also provides the more compact thermal barrier to the polymer chains. T_{max2} decreases for the incorporation of kaolin due to the loss of water around 450°C, while feldspar does not decompose at this temperature, so no visible changes were noticed in feldspar-incorporated RPUF.[36,37] For the foams with 10% concentration of the mineral filler, a degradation peak has been observed around 200°C, which is attributed to the presence of unreacted raw materials owing to an incomplete reaction of the polyol with the polyisocyanate and the deterioration in the structure of the RPUF. Amount of residue produced was also found to be increased on increasing the filler concentration in RPUF, which supports early extinguish of the flame in real fire condition.[38]

TABLE 3 Thermal characteristics of RPUF incorporated with different mineral fillers

Concentration of Filler	$T_{5\%}$ (°C) ± 2	T_{max1} (°C) ± 2	T_{max2} (°C) ± 2	Residual Mass (%) ± 0.1
0%	192	330	488	8.1
2% kaolinite clay	220	344	467	9.5
4% kaolinite clay	233	335	474	10.9
6% kaolinite clay	260	339	476	15.0
8% kaolinite clay	224	341	477	17.3
10% kaolinite clay	196	335	475	13.1
2% feldspar	202	339	480	15.5
4% feldspar	212	342	478	16.9
6% feldspar	212	337	475	18.1
8% feldspar	218	337	470	18.9
10% feldspar	190	332	473	17.1

3.4 | Flammability testing

Figures 9 and 10 show the fire behavior of RPUF composites containing varied concentration of mineral fillers, and relative data are summarized in Table 4.

The cone calorimetry provides quantitative analysis of the fire behavior of foams. HRR is considered as an indication to measure the

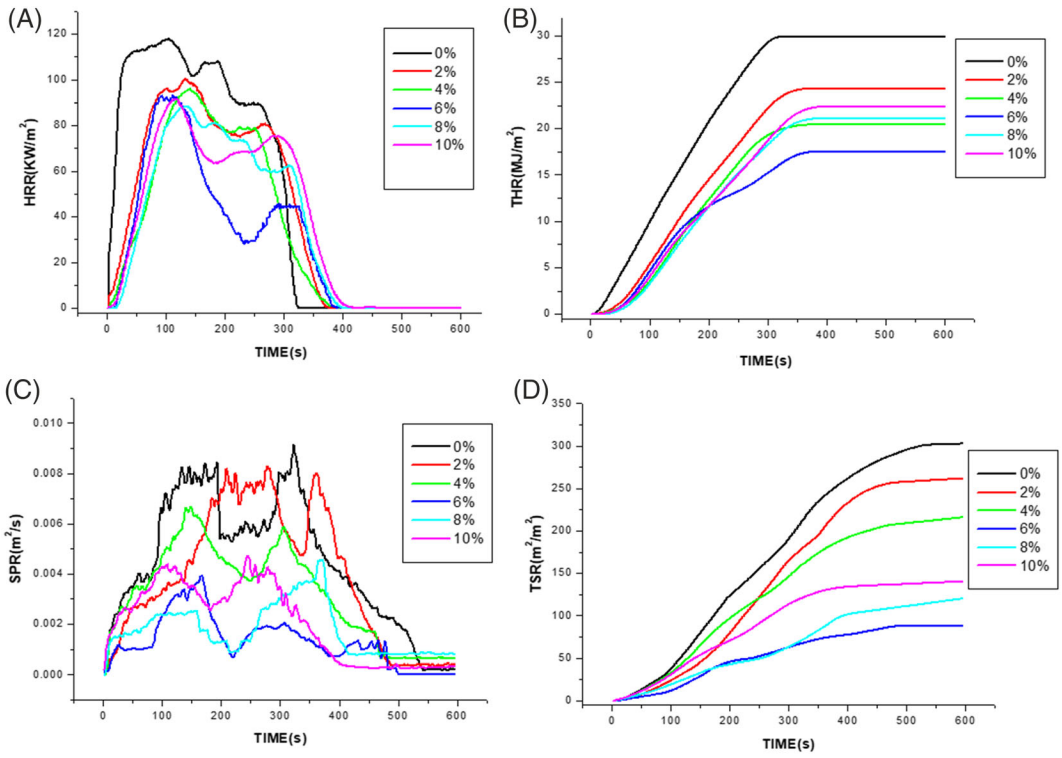


FIGURE 9 A, Heat release rate (HRR), B, total heat release (THR), C, smoke production rate (SPR), and D, total smoke release (TSR) of rigid polyurethane foams (RPUFs) incorporated with kaolinite clay

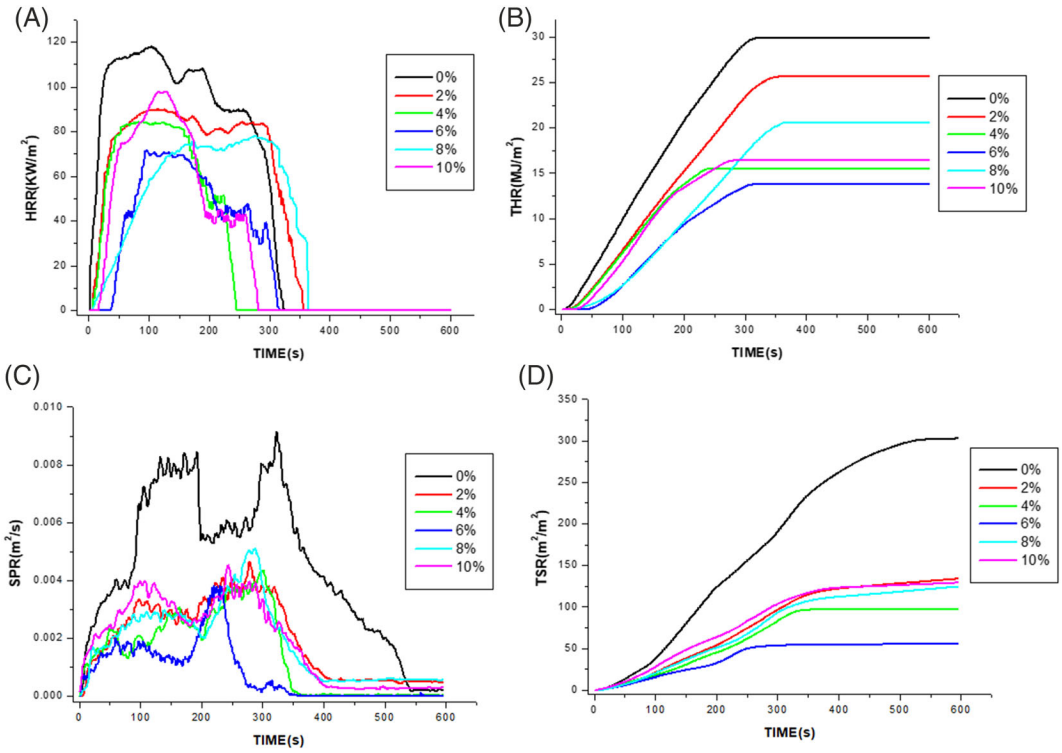


FIGURE 10 A, Heat release rate (HRR), B, total heat release (THR), C, smoke production rate (SPR), and D, total smoke release (TSR) of rigid polyurethane foams (RPUFs) incorporated with feldspar

intensity of the flame. The polyurethane foams exhibited two apparent HRR peaks: The first peak is attributed to the surface pyrolysis of foam and release of flammable low molecular weight products such as primary or secondary amine, isocyanate, alcohol, and olefin, while the second one is probably due to the degradation of previously formed charred protective layer and the release of entrapped flammable

TABLE 4 Flammability characteristics of RPUF incorporated with mineral fillers

Concentration of Filler	PHRR (kW/m^2) ± 5	THR (MJ/m^2) ± 0.5	TSR (m^2/m^2) ± 10	h_{eff} (MJ/kg) ± 5%	CO (kg/kg) ± 0.002	CO_2 (kg/kg) ± 0.05	CO/CO_2 ratio ± 0.003	Residue (%) ± 0.5
0%	118	29.8	347	9.26	0.078	1.95	0.040	11.5
2% kaolinite clay	101	24.3	335	6.63	0.058	1.51	0.038	15.6
4% kaolinite clay	96	20.4	216	5.41	0.042	1.25	0.033	17.9
6% kaolinite clay	93	17.5	88	4.22	0.032	1.10	0.029	18.1
8% kaolinite clay	89	21.1	120	4.91	0.046	1.30	0.035	22.2
10% kaolinite clay	92	22.4	140	4.84	0.065	1.52	0.041	23.5
2% feldspar	90	25.6	134	6.84	0.051	1.32	0.038	15.6
4% feldspar	84	15.5	97	4.01	0.038	1.15	0.033	16.7
6% feldspar	72	13.7	55	3.48	0.028	0.98	0.028	18.2
8% feldspar	78	20.6	124	4.95	0.032	1.05	0.030	20.8
10% feldspar	98	16.4	129	3.78	0.055	1.45	0.037	22.2

Abbreviations: PHRR, peak heat release rate; THR, total heat release; TSR, total smoke release.

gases.[39] The peak heat release rate (PHRR) decreased from 118 ± 5 kW/m^2 to 89 ± 5 kW/m^2 and 72 ± 5 kW/m^2 on the incorporation of 8% kaolinite clay and 6% feldspar, respectively, in comparison with the neat RPUF (Figures 9A and 10A). This behavior is probably due to the formation of a distended charred layer in the mineral filler–incorporated foams, as the protective char forms a barrier for the heat and mass transfer.[40] Furthermore, mineral fillers decompose endothermically and release some nonflammable gases, which decreases the rate of heat release.[12] Figures 9B and 10B showed that neat RPUF exhibited greater THR than foams with mineral fillers and 6% feldspar-incorporated foam showed minimum THR among all the samples. This indicates that mineral fillers prevent the heat transfer, consequently protecting the material beneath the charred layer from further burning. Formation of residue is another important parameter to quantify the fire properties. Incorporation of mineral fillers increased the amount of residue from 11.5 ± 0.5% to 23.5 ± 0.5% and 22.2 ± 0.5% for RPUF containing maximum filler concentration of 10% kaolinite clay and 10% feldspar, respectively. The higher residue for the filler-incorporated foams is probably due to the formation of more residue from the endothermal decomposition of the mineral fillers.[12]

Foam toxicity can be measured by the release of the carbon monoxide during the fire condition. So toxicity can be demonstrated by the ratio of CO/CO_2, which attributes to the extent of complete combustion of the foam. Higher the value of CO/CO_2 ratio, higher is the extent of incomplete combustion, consequently generating more toxic smoke. Incorporation of the mineral fillers decreases the release of toxic gases such as carbon monoxide and carbon dioxide.

Effective heat of combustion is determined by the ratio of THR to total mass loss. It was observed that the effective heat of combustion decreases on the increase in the filler content, which is attributed to the inhibition of flame or dilution of the gas phase by the incorporated fillers. Kaolinite clay and feldspar act as the thermal barrier due to the non-reactivity of the mineral fillers. Results show that 6% feldspar-incorporated RPUF demonstrated minimum effective heat of combustion, which is accredited to the optimized structure and properties. Feldspar starts to melt and decompose at 1150°C, while kaolinite clay dehydrates at a temperature around 450°C to give metakaolinite, which turns into mullite and amorphous silica on the heating above 1000°C. It supports the high temperature performance of mineral fillers and reduces the flammability of RPUF.[36,37] The smoke production is another important parameter to evaluate the fire behavior of any material, which is determined by SPR and total smoke release (TSR). Figures 9C and 10C show the trends in smoke production, which exhibits similar trend as shown by HRR. The SPR corresponding to the first peak decreased from 0.008 to 0.004 m^2/s and 0.003 m^2/s for foams with 6% kaolinite clay and 8% feldspar, respectively. The TSR of these foams was decreased to 88 ± 10 m^2/m^2 and 55 ± 10 m^2/m^2 for 6% feldspar and 6% kaolin, respectively, in comparison with 347 ± 10 m^2/m^2 for the neat foam (Figures 9D and 10D). This behavior of RPUFs attributes to the denser structure of the charred layer that restricts the spread of flame and diluted the gas phase. The smoke suppression is attributed to the generation of the charred protective layer that hinders the release of the smoke.

As the mineral fillers are noncombustible fillers, incorporation of these fillers reduces the flammability by reducing the total amount of flammable materials and hindering the rate of diffusion of oxygen. The performance of the filler in flame retardancy is highly dependent on the compatibility of the filler and heat absorbed by it during decomposition. Owing to the endothermic decomposition of the kaolinite clay, heat required to vaporize the same amount of fuel is increased; additionally, existence of gas phase flame diluents (eg, water) also tend to reduce the temperature of the flame, consequently reducing the fraction of heat transmitting to the foam. A critical concentration of free radicals is required for the self-sustaining flame. So flame can be extinguished by reducing the concentration of free radicals, which is performed by the release of water in decomposition of kaolinite clay.[12,36]

Feldspar is deposited as inert layer on the surface of the decomposing RPUF due to its noncombustible behavior. It prevents

the incoming radiation and oxygen reaching the decomposing RPUF and restricts the release of the flammable pyrolysis products and smoke to the gas phase.[37]

4 | CONCLUSION

RPUFs were prepared by incorporating different mineral fillers to enhance the mechanical, thermal, and flame retardant properties of the foams. The fillers considered in the present investigation exhibited considerable improvement in the properties of the RPUFs. Experimental results showed that RPUF incorporated with 8% feldspar demonstrates maximum compressive strength of 4.98 MPa, which is 182% greater than that of unreinforced RPUF. Thermal stability and flame retardancy of foams were also found to be improved on the addition of the mineral fillers. RPUF incorporated with clay demonstrated best thermal properties and showed increase in 5% weight loss temperature from 192°C to 260°C-270°C. Char residue was also found to be increased from 8 to 16 wt% on the incorporation of mineral fillers. Cone calorimeter test of RPUF with varied concentration of mineral fillers demonstrated a fair enhancement in the flame retardancy of RPUF, and among all the samples, best properties were shown by the foams with 6% concentration of any of the mineral filler. Of these fillers, feldspar showed higher flame retardancy; it suggests that feldspar has better fire performance in the mineral filler studied in this research; 8% feldspar shows best mechanical properties, but thermal properties as well as fire performance start to deteriorate on the incorporation of 8% feldspar. This behavior is accredited to the fact that agglomeration of the filler commences at 8% concentration of filler but the cellular structure of the foam is not effected considerably as shown in the SEM, so thermal and fire performance starts to deteriorate due to lesser barrier effect of the filler. While higher concentration of filler in the cell walls of RPUF strengthens the structure, consequently increasing its compressive strength. A set of data is said to be reproducible if it falls under the probability of 95%. The findings of present study show up to 5% uncertainty, so results are considered as reproducible and repeatable. As the substitution of petroleum-based feedstock with agricultural oil, for the production of polymers, is a potential field of research, and improvement of their properties is the need of the hour. This study shows that the incorporation of mineral fillers improves the mechanical, thermal, and flame retardant properties of vegetable oil–based RPUFs, so mineral fillers may be utilized as a unique flame retardant with good mechanical and thermal properties. These mineral fillers–incorporated RPUFs may be used for fire safety applications in buildings, for carving models and prototypes.

ACKNOWLEDGEMENT

This work was supported by Defence Research and Development Organisation (DRDO), India (grant number: CC/TM/ERIPR/GIA/15-16/029).

ORCID

Raminder Kaur https://orcid.org/0000-0002-9401-8337

REFERENCES

1. Gama NV, Soares B, Freire CSR, et al. Bio-based polyurethane foams toward applications beyond thermal insulation. *Mater des*. 2015;76:77-85. https://doi.org/10.1016/j.matdes.2015.03.032
2. https://www.generalplastics.com/technical-papers/dielectric-materials-use-radomes. https://www.generalplastics.com/technical-papers/dielectric-materials-use-radomes. Accessed November 13, 2017.
3. Rigid polyurethane foam sheets—high density hard foam blocks|general plastics—general plastics. https://www.generalplastics.com/rigid-foams. Accessed June 29, 2018.
4. Miao S, Wang P, Su Z, Zhang S. Vegetable-oil-based polymers as future polymeric biomaterials. *Acta Biomater*. 2014;10(4):1692-1704. https://doi.org/10.1016/j.actbio.2013.08.040
5. Ribeiro Da Silva V, Mosiewicki MA, Yoshida MI, Coelho Da Silva M, Stefani PM, Marcovich NE. Polyurethane foams based on modified tung oil and reinforced with rice husk ash I: synthesis and physical chemical characterization. *Polym Test*. 2013;32(2):438-445. https://doi.org/10.1016/j.polymertesting.2013.01.002
6. Piszczyk L, Strankowski M, Danowska M, Haponiuk JT, Gazda M. Preparation and characterization of rigid polyurethane-polyglycerol nanocomposite foams. *Eur Polym J*. 2012;48(10):1726-1733. https://doi.org/10.1016/j.eurpolymj.2012.07.001
7. Piszczyk L, Strankowski M, Danowska M, Hejna A, Haponiuk JT. Rigid polyurethane foams from a polyglycerol-based polyol. *Eur Polym J*. 2014;57:143-150. https://doi.org/10.1016/j.eurpolymj.2014.05.012
8. Agrawal A, Kaur R, Walia RS. PU foam derived from renewable sources: perspective on properties enhancement: an overview. *Eur Polym J*. 2017;95:255-274.
9. Serban DA, Weissenborn O, Geller S, Marsavina L, Gude M. Evaluation of the mechanical and morphological properties of long fibre reinforced polyurethane rigid foams. *Polym Test*. 2016;49:121-127. https://doi.org/10.1016/j.polymertesting.2015.11.007
10. Xue B, Wen J, Sun R. Lignin-based rigid polyurethane foam reinforced with pulp fiber: synthesis and characterization. *ACS Sustain Chem Eng*. 2014;2(6):1474-1480.
11. Pukánszky B. Mineral filled polymers. In: *Reference Module in Materials Science and Materials Engineering*. Elsevier; 2016. doi:https://doi.org/10.1016/B978-0-12-803581-8.02598-4
12. Hull TR, Witkowski A, Hollingbery L. Fire retardant action of mineral fillers. *Polym Degrad Stab*. 2011;96(8):1462-1469. https://doi.org/10.1016/j.polymdegradstab.2011.05.006
13. Gao X, Zhou B, Guo Y, et al. Synthesis and characterization of well-dispersed polyurethane/$CaCO_3$ nanocomposites. *Colloids Surfaces A Physicochem Eng Asp*. 2010;371(1–3):1-7. https://doi.org/10.1016/j.colsurfa.2010.08.036
14. Alavi Nikje MM, Garmarudi AB, Haghshenas M. Effect of talc filler on physical properties of polyurethane rigid foams. *Polym Plast Technol Eng*. 2006;45(11):1213-1217. https://doi.org/10.1080/03602550600887541
15. Fan H, Tekeei A, Suppes GJ, Hsieh FH. Properties of biobased rigid polyurethane foams reinforced with fillers: microspheres and nanoclay. *Int J Polym Sci*. 2012;2012:1-8. https://doi.org/10.1155/2012/474803
16. Estravís S, Tirado-Mediavilla J, Santiago-Calvo M, Ruiz-Herrero JL, Villafañe F, Rodríguez-Pérez MÁ. Rigid polyurethane foams with infused nanoclays: relationship between cellular structure and thermal

conductivity. *Eur Polym J*. 2016;80:1-15. https://doi.org/10.1016/j.eurpolymj.2016.04.026

17. Czupryński B, Paciorek-Sadowska J, Liszkowska J. Properties of rigid polyurethane-polyisocyanurate foams modified with the selected fillers. *J Appl Polym Sci*. 2010;115:2460-2469.
18. Nik Pauzi NNP, Majid RA, Dzulkifli MH, Yahya MY. Development of rigid bio-based polyurethane foam reinforced with nanoclay. *Compos Part B Eng*. 2014;67:521-526. https://doi.org/10.1016/j.compositesb.2014.08.004
19. Zheng X, Wang G, Xu W. Roles of organically-modified montmorillonite and phosphorous flame retardant during the combustion of rigid polyurethane foam. *Polym Degrad Stab*. 2014;101(1):32-39. https://doi.org/10.1016/j.polymdegradstab.2014.01.015
20. Zieleniewska M, Leszczyński MK, Szczepkowski L, et al. Development and applicational evaluation of the rigid polyurethane foam composites with egg shell waste. *Polym Degrad Stab*. 2015;132:78-86. https://doi.org/10.1016/j.polymdegradstab.2016.02.030
21. Zhu M, Bandyopadhyay-Ghosh S, Khazabi M, Cai H, Correa C, Sain M. Reinforcement of soy polyol-based rigid polyurethane foams by cellulose microfibers and nanoclays. *J Appl Polym Sci*. 2012;124:4702-4710.
22. Atabek LS. The flame-retardant effect of calcium hypophosphite in various thermoplastic polymers. 2019;(January):294–302. https://doi.org/10.1002/fam.2700
23. Chen XY, Fan XY, Wang Z. Synergistic effect of carbon and phosphorus flame retardants in rigid polyurethane foams. October 2017;2018(4):447-453. https://doi.org/10.1002/fam.2511
24. Wang X, Zhang P, Huang Z, Xing W, Song L, Hu Y. Effect of aluminum diethylphosphinate on the thermal stability and fl ame retardancy of flexible polyurethane foams. *Fire Saf J*. 2019;106(April):72-79. https://doi.org/10.1016/j.firesaf.2019.04.001
25. Wang S, Du X, Jiang Y, et al. Journal of colloid and interface science synergetic enhancement of mechanical and fire-resistance performance of waterborne polyurethane by introducing two kinds of phosphorus–nitrogen flame retardant. *J Colloid Interface Sci*. 2019;537:197-205. https://doi.org/10.1016/j.jcis.2018.11.003
26. Zhu H, Xu S. Synthesis and properties of rigid polyurethane foams synthesized from modified urea-formaldehyde resin. *Construct Build Mater*. 2019;202:718-726. https://doi.org/10.1016/j.conbuildmat.2019.01.035
27. Feldspar|Minerals Education Coalition. https://mineralseducationcoalition.org/minerals-database/feldspar/. Accessed September 19, 2017.
28. What is kaolin clay? Benefits, powder, skin, side effects, properties of China/white clay. https://durablehealth.net/kaolin-clay/kaolin-clay-benefits-side-effects-properties/. Accessed September 18, 2017.
29. Pruett RJ. Kaolin deposits and their uses: Northern Brazil and Georgia, USA. *Appl Clay Sci*. 2016;131:3-13. https://doi.org/10.1016/j.clay.2016.01.048
30. Kumar M, Kaur R. Effect of different formulations of MDI on rigid polyurethane foams based on castor oil. *Int J Sci Res Rev*. 2013;2:29-42.
31. Kaur R, Kumar M. Function of silicon oil in the castor oil based rigid polyurethane foams. *J Polym Eng*. 2013;33(9):875-880. https://doi.org/10.1515/polyeng-2013-0083
32. Li X, Tabil LG, Panigrahi S. Chemical treatments of natural fiber for use in natural fiber-reinforced composites: a review. *J Polym Environ*. 2007;15(1):25-33. https://doi.org/10.1007/s10924-006-0042-3
33. Ciecierska E, Jurczyk-Kowalska M, Bazarnik P, et al. Flammability, mechanical properties and structure of rigid polyurethane foams with different types of carbon reinforcing materials. *Compos Struct*. 2016;140:67-76.
34. Malik M, Kaur R. Mechanical and thermal properties of castor oil-based polyurethane adhesive: effect of TiO_2 filler. *Adv Polym Technol*. 2016;(1):1-7. https://doi.org/10.1002/adv.21637
35. Zhang M, Luo Z, Zhang J, Chen S, Zhou Y. Effects of a novel phosphorus-nitrogen flame retardant on rosin-based rigid polyurethane foams. *Polym Degrad Stab*. 2015;120:427-434. https://doi.org/10.1016/j.polymdegradstab.2015.08.001
36. Ptáček P, Šoukal F, Opravil T, Havlica J, Brandštetr J. The kinetic analysis of the thermal decomposition of kaolinite by DTG technique. *Powder Technol*. 2011;208(1):20-25. https://doi.org/10.1016/j.powtec.2010.11.035
37. Zhong Y, Gao J, Chen P, Guo Z. Recovery of potassium from K-feldspar by thermal decomposition with flue gas desulfurization gypsum and $CaCO_3$: analysis of mechanism and kinetics. 2016:2–10. https://doi.org/10.1021/acs.energyfuels.6b01915
38. França de Sá S, Ferreira JL, Pombo Cardoso I, Macedo R, Ramos AM. Shedding new light on polyurethane degradation: assessing foams condition in design objects. *Polym Degrad Stab*. 2017;144:354-365. https://doi.org/10.1016/j.polymdegradstab.2017.08.028
39. Yang R, Wang B, Han X, Ma B, Li J. Synthesis and characterization of flame retardant rigid polyurethane foam based on a reactive flame retardant containing phosphazene and cyclophosphonate. *Polym Degrad Stab*. 2017;144:62-69. https://doi.org/10.1016/j.polymdegradstab.2017.08.008
40. Zhang M, Zhang J, Chen S, Zhou Y. Synthesis and fire properties of rigid polyurethane foams made from a polyol derived from melamine and cardanol. *Polym Degrad Stab*. 2014;110:27-34. https://doi.org/10.1016/j.polymdegradstab.2014.08.009

How to cite this article: Agrawal A, Kaur R, Walia RS. Investigation on flammability of rigid polyurethane foam-mineral fillers composite. *Fire and Materials*. 2019;1–11. https://doi.org/10.1002/fam.2751

Investigations of tribodynamic characteristics of chrome steel pin against plain and textured surface cast iron discs in lubricated conditions

Roop Lal and R.C. Singh
Department of Mechanical Engineering, Delhi Technological University, Delhi, India

Abstract

Purpose – The purpose of this paper was to carry out an experimental study to investigate tribodynamics at the interface of chrome steel pin with the plain and textured surface cast iron discs in a fully flooded condition.

Design/methodology/approach – A friction and wear tester (tribometer) was used for conducting the experiment at different operating parameters. Lubricating oil used in this experiment was SAE-15W40, which directly supplied at the interface and ensured a fully flooded lubricating condition. Measurements of different parameters, such as friction coefficient, the rate of wear, the amplitude of vibration, rise in temperature and intensity of sound, were considered at different loads. The input parameters were taken in a range viz. Loads from 1.9 MPa to 3.1 MPa, sliding speeds from 2 to 10 m/s and distance travelled ranged from 1000 to 5000 m.

Findings – In the investigation, it was found that there was a decrease in friction coefficient and wear rate at the interface formed by the chrome steel pin with textured surface cast iron disc when compared with friction coefficient and wear rate at the interface formed by the chrome steel pin and the plain surface disc of cast iron.

Originality/value – Also, it was inferred that temperature, vibration and noise level, at the interface formed by the chrome steel pin and plain surface cast iron disc, increased, which governed the interface stability. The novelty of this investigation lies in the identification of the parameters on which interface stability depends.

Keywords Friction, Wear, Pin-on-disc, Vibration, Lubrication, Textured discs

Paper type Research paper

1. Introduction

Hydrodynamic journal bearings are widely used in supporting and guiding shafts efficiently. It also possesses inherent high damping capacity, which is beneficial from the perspective of dynamics. Through the progress of research, it had been identified that thin film lubrication is beneficial in journal bearings. Hence, research on journal bearing operating in mixed film lubrication regime is essential for reduction in associated friction and wear. Different techniques are used by researchers across the globe for improving the performance behavior of the journal bearing operating under thin film lubrication. A brief literature review is presented in the following paragraphs for understanding the current status of research in the field of thin film lubrication in bearings.

Authors investigated aspects such as the effect of hybrid texturing, laser texturing on steel surface filled with poly phenylene sulphide (PPS) powder, ball-on-disk sliding tests performed under dry lubrication conditions and pin-on-disc model with two diesel fuels, concluding that hybrid texturing protects metal surfaces with the formation of a polymer transfer layer. The dependency of the surface texture parameters on the friction coefficient to redefine surface topography of a product after the surface finishing processes was investigated, which reported a dip in friction coefficient with an increase in load (Cho, 2016; Kuilenburg *et al.*, 2012; Djoufack *et al.*, 2015). Authors evaluated the dependency of partial laser surface texturing (LST) on the reduction of friction in piston rings and concluded that partial LST cylindrical piston rings for variable engine speeds under near-half-load conditions reduce friction by 25 per cent in comparison to non-textured conventional rings, vibrations, fuel and lubricating oil consumption (Etsion and Sher, 2009; Ryk and Etsion, 2006; Sudeep *et al.*, 2013; Pandey *et al.*, 2013). Researchers also studied the tribological performance of micro dimples, which formed the base of frictional tests, laser peen textured copper surface in starved lubrication and multi-shape textured surface produced by laser ablation; they concluded that combination of geometric pattern surface helped in reducing friction significantly in both dry and lubricated conditions (Kim *et al.*, 2014; Li *et al.*, 2014; Zenebe Segu and Hwang, 2015). Investigation of the effect of machining environment on the quality of machining under dry and wet conditions yield a reduction in friction and wear values

The current issue and full text archive of this journal is available on Emerald Insight at: **www.emeraldinsight.com/1708-5284.htm**

World Journal of Engineering
© Emerald Publishing Limited [ISSN 1708-5284]
[DOI 10.1108/WJE-09-2018-0311]

The authors would like to offer special thanks to the Nano Research Facility of Indian Institute Technology, Delhi, India. They also extend special thanks to KOM, DOM and Tribology Lab, DTU, Delhi, India, for their kind help in conducting the practical session.

Received 18 September 2018
Revised 10 February 2019
Accepted 20 February 2019

at higher sliding speed and heavy loads (Hu and Xu, 2016). Wear behavior of textured surfaces specimen, brake disc material with groove textured, and un-textured counter-specimen, both specimen textured and un-textured with varying spacing under dry condition, simultaneous variations in dimple densities for unidirectional sliding under starved lubrication, by varying the normal loads were considered to evaluate the behavior at room and elevated the temperature. The results showed reduced values of friction and wear indicating the small texture spacing helps in achieving the highest friction coefficient and the shallowest wear depth, which are actually the combined effect of increased surface roughness and the micro-cutting effect by the texture edges which reduced real contact area, and led to entrapment of wear debris (Wos *et al.*, 2017, 2018; Mo *et al.*, 2013; Xing *et al.*, 2017). Authors (Olofinjana *et al.*, 2015) investigated the influence of laser surface texturing on the formation and durability of tribo chemical films. Vladescu *et al.*(2016) and Gachot *et al.* (2017) analyzed the tribological behavior under synthetic lubrication with and without additives. The fused silica specimens with textures were studied against a convex steel pad. The fabrication techniques of surface texturing and their effect on the tribological behavior of materials and textured gray cast iron (CI) using lubricated block-on-ring tests were studied by da Silva and Costa (2017). The results showed improvement in the relative performance of the textured specimens due to the development of progressive contact into the mixed and boundary regime in most of the cases.

Many researchers used a numerical model to investigate the tribological behavior and deduced an optimal design of a multi-dimpled surface. The laser fabrication technique was used for developing texture on stainless steel and low alloy carbon steel. The influence of machining parameters and vibration modes on the tribological behavior along with the tribological properties were investigated considering metal-on-metal and metal-on-plastic contacts. Results showed extremely high static friction coefficients, reduction in dynamic friction coefficient reduced wear rate and adhesion between pin and samples surface. The dimpled specimen yielded a further reduced friction coefficients and wear rates (Wei *et al.*, 2013; Dunn *et al.*, 2015; Amini *et al.*, 2016; Jones and Schmid, 2016). Textured the specimen by whirling electrical discharge texturing (WEDT). The friction tests yielded to decrease in the friction coefficient under the boundary lubrication and the mixed lubrication condition (Yamaguchi *et al.*, 2016). The composite of PEK and PBI was studied by authors Bijwe *et al.* (2015). Textures formed by honing and plateau honing process on the surfaces of the cylinder liners. Specimens were cut from gray CI cylinder liners. Then experiments were performed under lubrication conditions on an oscillating wear tester. It showed excellent tribo-potential, which has a coefficient of friction of level 0.5 and the specific wear rate 10^{-16} m^3/Nm (Grabon *et al.*, 2017). Wear resistance and friction coefficient of non-smooth surfaces were investigated. Effect of massive laser shock peening (LSP) on dry sliding wear were evaluated and discussed (Luo *et al.*, 2014). Authors discussed microscale textures engraved on surfaces of carbides with the help of a laser. These textures were filled with a solid lubricant to act as coatings on the surfaces of cemented carbide. Then dry lubrication sliding tests were carried out against austenitic steel balls in the atmospheric air. The results showed lower friction in case of micro scale textured cemented carbides, which compared with un-textured cemented carbides (Meng *et al.*, 2018). Double-layer thermal barrier coatings were formed on SS310 substrates with the help of high-velocity oxygen fuel thermal spray method. Coatings were exposed to a high-temperature corrosive environment and then tested for performance. The result reported that there was an improvement in the resistance to corrosion (Taie *et al.*, 2018). Authors also presented a high-temperature method of indentation. This method was developed for the investigation of residual stress and high-temperature fracture toughness in a typical TBC. Nanostructured 8 Wt.% yttrium partially stabilized zirconia (YSZ) coating. It was observed that cracks were caused by the high-temperature indentation tests (Qu *et al.*, 2018). Authors prepared aluminum melt fly ash particles and studied their effect on the wear (Sharma *et al.*, 2017). This study reported the synthesis of Mg_2Si co-doped with Bi and Sb and carried out spark plasma reaction sintering, shown enhancement in mechanical properties. (Muthiah *et al.*, 2017) Tribological performance analysis was performed of textured steel surfaces and plain CI under lubricating and dry conditions (Singh *et al.*, 2016; Lal and Singh, 2018; Rana *et al.* (2018)).

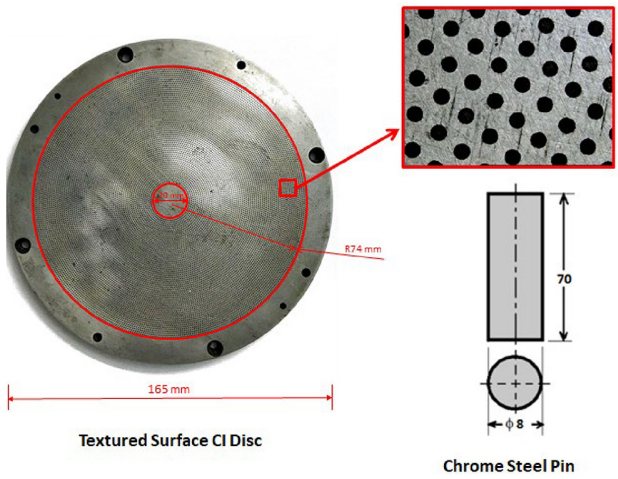
From the above literature review, it was found that a reduction in wear and friction in bearings is a vital task. It also found that there were few studies available for exploring the behaviors of the journal bearing in oil-lubricated conditions. Therefore, the objective of this research paper is to experimentally find out tribodynamic performances of interfaces formed by chrome steel pin with plain surface and textured surface CI discs. Also, the study aims to further explore the possibility of manufacturing of textured surface CI journal and chrome-coated bearing.

2. Experimentation

2.1 Descriptions of discs

The specimens of plain surface CI and textured surface CI discs having diameter of 165 mm and thickness of 8 mm, shown in Figure 1, were manufactured through casting process for

Figure 1 Dimensional details plain surface and textured surface discs and pin



experimentation. Surfaces of both the discs were finished with the help of belt emery paper. The surface finish was measured by a surface-roughness-measuring instrument called telly surf. The measurement was taken at an interval of 45° in the radial direction and eight readings perpendicular to it. Obtained roughness value in terms of RMS was 0.15-0.3 μm. On one of the CI discs, the circular texture was formed by the chemical etching process. Stereo zoom microscope was used for measuring the diameter of the circular texture. The average diameter of circular texture was found to be 0.5965 mm ≈ 0.6 mm, and it was calculated that the 65 per cent disc surface was textured [Figure 7(a)]. Surface images of the plane surface CI disc taken with microscope are shown in Figure 7(b).

2.2 Descriptions of chrome steel pin

A cylindrical Pin of chrome steel had 8-mm diameter, and an axis of 70 mm was used in testing on specimens of plain surface and textured surface CI discs (Figure 1). The face of cylindrical pin leveled and debris removed from the face of the pin by rubbing it on the soft paper. The surface plate was used to check the flatness of the pin face. The part of pin length was held rigidly in the clamp, and 4 mm of the pin was protruding out of the fixture. This ensured that during the test, the part of the pin outside the clamp did not bend.

2.3 Description of oil used in lubrication

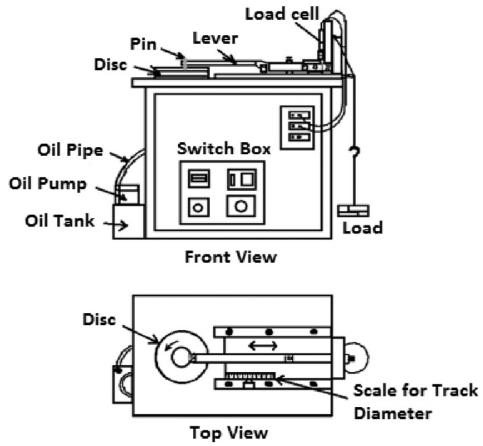
Lubricating oil used was 15W40. This oil was considered for the experiment because it is generally used in our tropical regions. The characteristics of this lubricating oil were determined at 40°C in the laboratory. They were:

- as per ASTM D – 445, kinematic viscosity = 160.32 mm^2/sec;
- as per ASTM D – 4052, density = 0.8744 g/cm^3; and
- as per ASTM D – 287, specific gravity = 0.8752.

2.4 Description of friction and wear test rig

Figure 2 shows the diagram of wear and friction tester used for this experimental study according to ASTM G-99. It had a motor having different speeds depending upon the position of the rheostat. Circular disc specimen was fixed at four points on the pitch circle on the horizontal fixture. It had a sensor for speed and the probe for wear measurement. A pan having a horizontal platform was used for placing different weights. This a horizontal platform support weights in the gravity direction. The applied weight/force was transmitted to the unilever via wire, pulleys and a bell-crank mechanism arrangement. A probe was fixed on the side of the bell crank lever which detected/determined force of friction. The rig has a speed range of 200-2,000 rpm as per the specifications and could measure friction force up to 200 N, with 0.1 N as the least count. Frictional force, Newton (N), can be measured with an accuracy of 2 per cent.

Figure 2 Wear and friction tester (tribometer)



Specifications of materials used in these experiments were:

- Chrome steel pin having a diameter of 8 mm; Discs:
- A CI disc had a plain surface. Diameter = 165 mm and thickness 8 mm
- Another CI disc had textured surface, Diameter = 165 mm and thickness 8 mm.

Experiments were planned in such a way that each factor had been analyzed individually. The time required for experimenting was significantly reduced. This plan was efficient for investigating the effect of individual factors, as well as investigation of the effects of multiple performance factors. The experiments were conducted at Tribology Lab, DTU, Delhi. The number of experiments performed on each disc was 27. The tribological performance studies were undertaken at different loads – loads taken in terms of pressure varying from 1.95 to 3.1 MPa, sliding speeds taken between 2 and 10 m/s, which is equal to 200-2,000 rpm. Also, distance traveled by the pin over the disc was between 1,000 and 5000 m. (Figure 3)

2.5 Elemental mapping analysis of the material of discs

2.6 EDS analysis results of the material of discs figure

3. Results and discussions

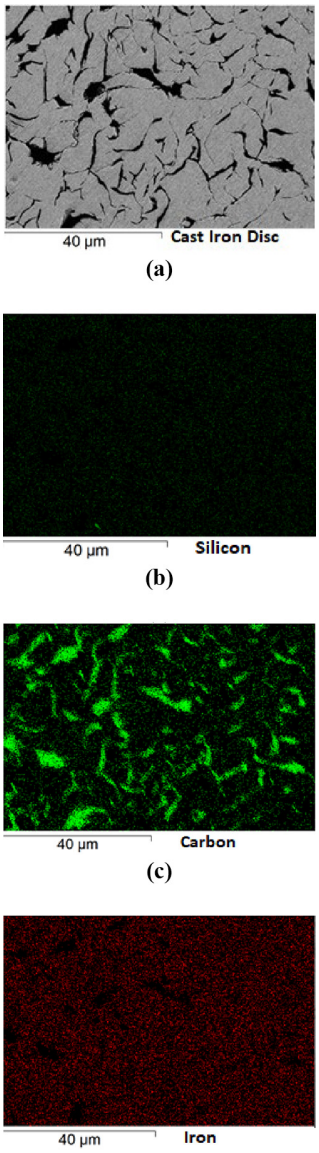
The common data for all experiments have been defined below along with the results and discussion:

1 The varying speed ranged between 1,000 and 5,000 rpm at various diameters of the disc. The value of loads applied in terms of the pressure were 1.95, 2.5 and 3.1 MPa.
2 The sliding speed was taken as 2, 6 and 10 m/s.
3 The pin used is made of chrome steel, with the dimensions 8-mm diameter and 70-mm-long axis.
4 The disc material has the following specifications:
   - plain surface CI disc: diameter = 165 mm and thickness = 8 mm; and
   - textured surface CI disc: diameter = 165 mm and thickness = 8 mm.
5 The lubricating oil used is 15W40 and weighing machine used has a least count of 0.0001 gm.

3.1 Friction coefficient

Graphs of friction coefficient (μ) vs contact pressure (1.95-3.1 MPa) with sliding speed 2, 6 and 10 m/s are plotted in Figure 8 for fully flooded lubrication conditions. A plain surface CI disc and a textured surface CI disc were taken for the determination of friction coefficient between discs and the

Figure 3 SEM image and EDAX spectra (a) SEM image of discs and (b, c, d, e, f) image of element mapping of the disc



Notes: (a) SEM image of discs and (b, c, d, e, f) image of element mapping of the disc

Figure 4 EDS spectrum of the material of discs

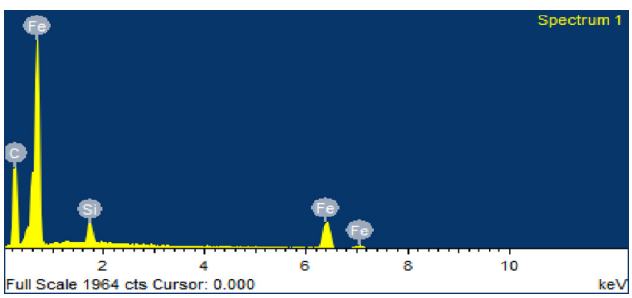
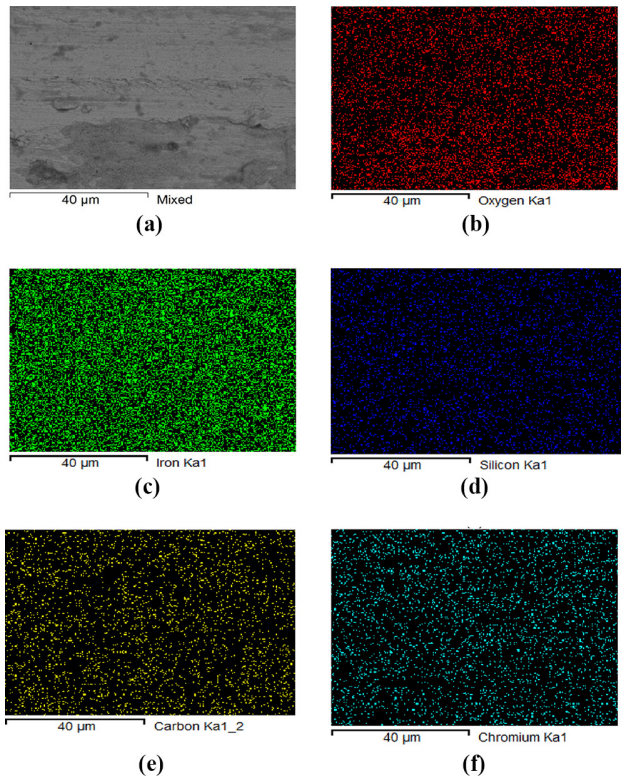
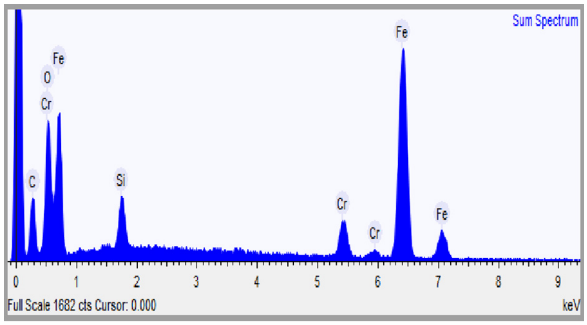


Figure 5 SEM image and EDAX spectra (a) SEM image of pin material and (b, c, d, e, f) image of element mapping of the pin sample



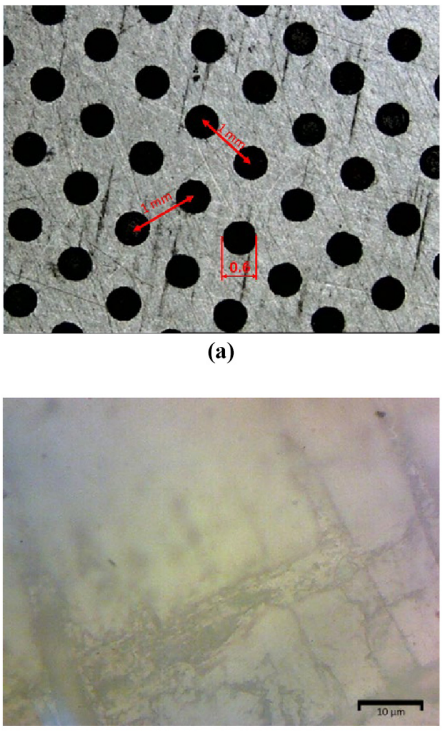
Notes: *Element maps: SEM images of pin material*: project: new; accumulator voltage: 15.0 kv; resolution: 256 × 192 pixels; resolution viewed: 100 per cent; process time: 5; image width: 82.6 µm

Figure 6 EDAX spectra of the pin sample



Notes: *Details of spectrum*: project: new; spectrum name: spectrum sum. *Acquisition conditions*: time of acquisition: 98.3 s; process time: 5 s; accelerating voltage: 15.0 (kV). *Quantification settings*: method of quantification, elements: normalized; coating: no coating. *Electron image*: width of image: 82.6 µm

Figure 7 (a) Surface images of textured surface CI disc with 35× zoom and (b) microscopic images of plain (smooth) surface of CI disc



(a)

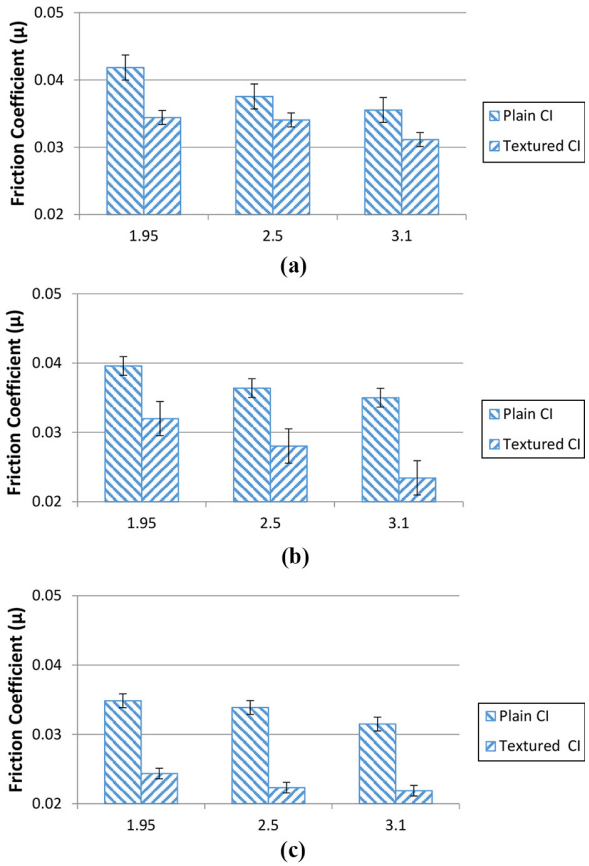
(b)

chrome steel pin. Better tribological parameters were obtained in case of textured surface CI disc in comparison to plain surface CI disc. Friction coefficient at the interface of the pin (chrome steel) and plain surface CI disc was observed to lie between 0.021 and 0.041. It was seen that a plain surface CI disc has a maximum value of friction coefficient as 0.041. While under similar conditions, the textured surface CI disc has a friction coefficient of 0.021-0.035. Also, the textured surface CI disc has a minimum friction coefficient. It was observed that the friction coefficient varies in the range of 0.021-0.041 in fully flooded conditions of lubrication developed by a continuous flow of lubricating oil on the disc. Also, it was ascertained that there was a continuous decrease in friction coefficient with an increase in load defined in terms of pressure.

3.2 Wear

Wear was measured through the weight loss method, i.e. reduction in weight of the pin. In Figure 9, the trend of wear is shown by wear characteristics, at the initial stage, where wear growth is fast. It might be due to the breaking of asperities in the contact surface of the tribopair. Variation of wear with different pressures is shown in the graphs displayed in Figure 9 for fully flooded lubrication. Wear decreases with increase in pressure, which may be due to elasto-hydrodynamic lubrication up to 3.1 MPa. Wear pattern on pin material increases initially but decreases later (Figure 9). It might be due to abrasion, followed by diffusion of surface debris. There exists a partial metal-to-metal contact between pin and discs. But because of the presence of a thin layer of lubricant at the interface, a continuous

Figure 8 Friction coefficient (μ) vs pressure



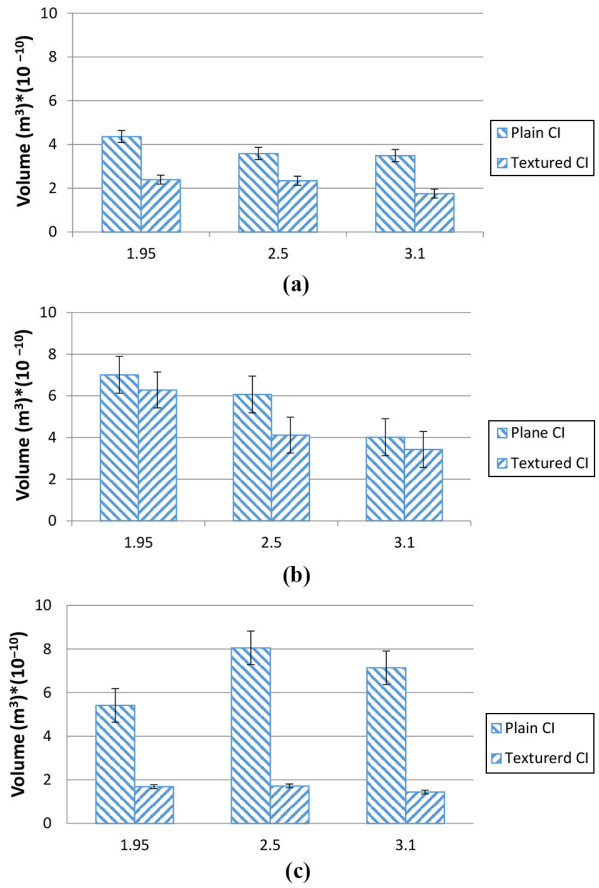
Notes: (a) Pressure (MPa) at speed constant: 2m/sec; (b) pressure (MPa) at speed constant: 6m/sec; (c) pressure (MPa) at speed constant: 10 m/sec

decrease in wear is observed. Also reduction in wear is observed at the interface of textured surface CI disc and chrome steel pin when fully flooded lubrication conditions are maintained. Overall wear values vary between 1.510^{-10} and $8\ 10^{-10}\ m^3$. It is inferred that the wear values were less at the interface of textured surface CI disc and steel pin due to the hydrodynamic lubrication effect of oil trapped in circular textures.

3.3 Vibrations amplitude

Experiments were performed at different speeds and loads taken in terms of contact pressure. Speeds were taken between 1,000 and 5,000 rpm along with the different predetermined track diameter of rotation on disc with varying the pressure. At these different conditions vibration amplitudes measured with the help of Vibro Expert Instrument and graphs which plotted as shown in Figure 10. It was observed that the overall vibration amplitude varies between 38 and 48 μm. When the interface was formed by textured surface CI disc and pin (chrome steel), the vibration amplitude varied from 38 to 45 μm under fully flooded lubrication conditions. Similarly, when the interface was formed by plain surface CI disc and pin under fully flooded lubrication conditions, the vibration amplitude varied between 45 and 48 μm. Reduction in the amplitude of vibration was

Figure 9 Volume (m3)*(10^{-10}) vs pressure (MPa)



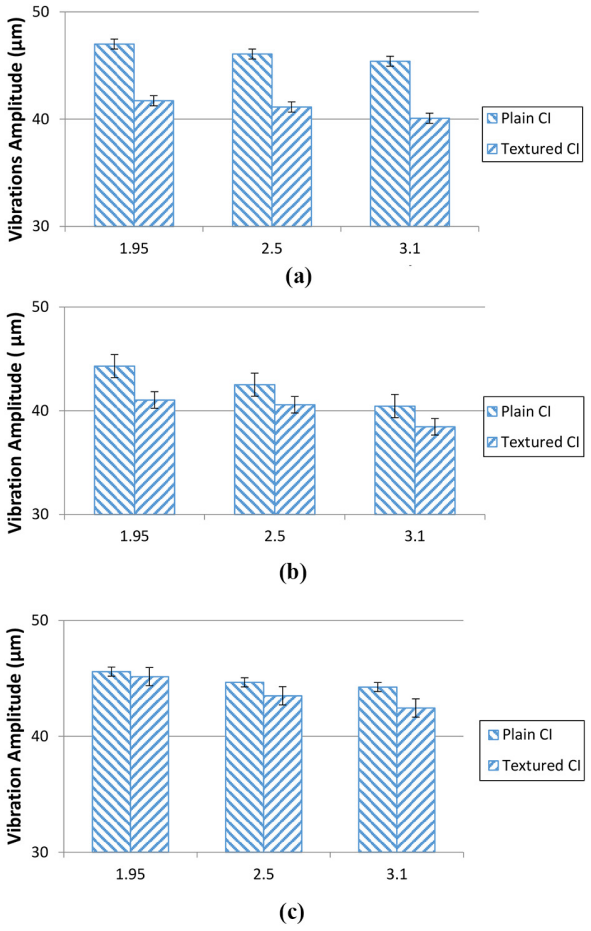
Notes: (a) Pressure (MPa) at speed constant: 2m/sec; (b) pressure (MPa) at speed constant: 6m/sec; (c) pressure (MPa) at speed constant: 10 m/sec

observed in the case of the interface formed by textured surface CI disc and pin due to energy absorption by the lubricant oil trapped in the textured grooves.

3.4 Sound

The noise pollution rules have been set acceptable at different noise limits as per its category of day or night. Noise pollution level limit for industries during daytime is 75 dB and during night time is 70 dB. Similarly, noise pollution level limit for residential areas during daytime is 55 dB and during night time is 45 dB. The fully flooded lubrication conditions were maintained by the continuous flow of the lubricating oil. But due to pressure at the interface of pin and discs, a thin layer of lubricating oil exists in the contact area, which develops a possibility of metal-to-metal contact between the discs and pin, leading to the creation of sound. Also, sound was produced because of the relative motion of disc with respect to the pin. The sound was measured by the sound meter, kept 25 cm away from wear and friction test rig. It was observed that the sound level varies between 66 and 73 dB (Figure 11). More noise levels were observed in case of plain surface CI disc as compared with textured CI discs, in the presence of thin layer lubrication conditions.

Figure 10 Vibration amplitude (μm) vs pressure (MPa)



Notes: (a) Pressure (MPa) at speed constant: 2m/sec; (b) pressure (MPa) at speed constant: 6m/sec; (c) pressure (MPa) at speed constant: 10 m/sec

3.5 Temperature rise

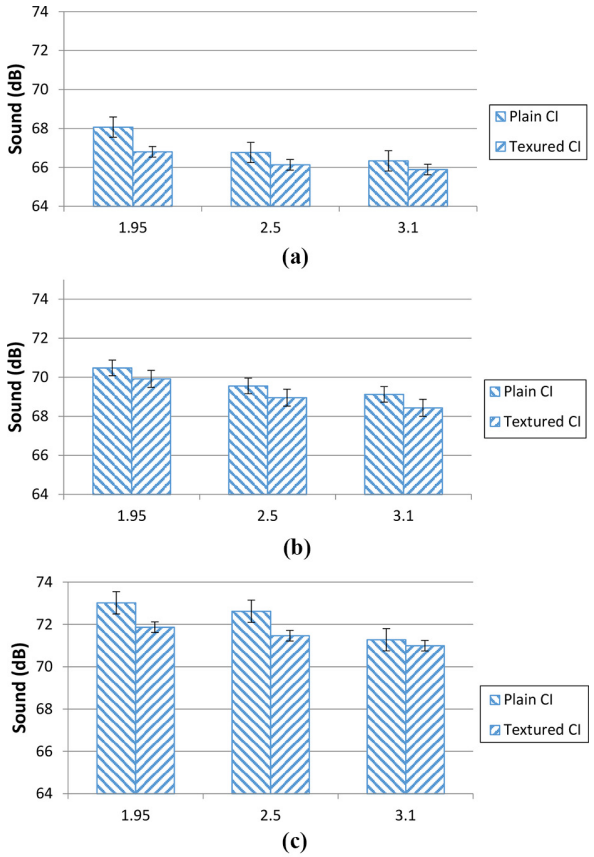
There was a nominal temperature rise due to the relative motion between discs and pin. Though fully flooded lubrication conditions were maintained by the continuous flow of lubricating oil, metal-to-metal contact existed because of a thin layer lubrication present at the contact area of the disc and pin. Also, temperature rise depends upon the roughness of the surfaces in contact and nature of the material along with the mechanical properties. Graphs between the rise in temperature at the interface of tribo-pairs formed by plain CI disc, textured surface CI disc with a pin (chrome steel) and pressure have been in shown Figure 12. It was found that temperature rise varies in the range of 4 and 6 K. The marginal rise in temperature was observed, which occurred because of a continuous flow of lubricating oil at the interface.

4. Conclusions

The following were observed during experimentation:

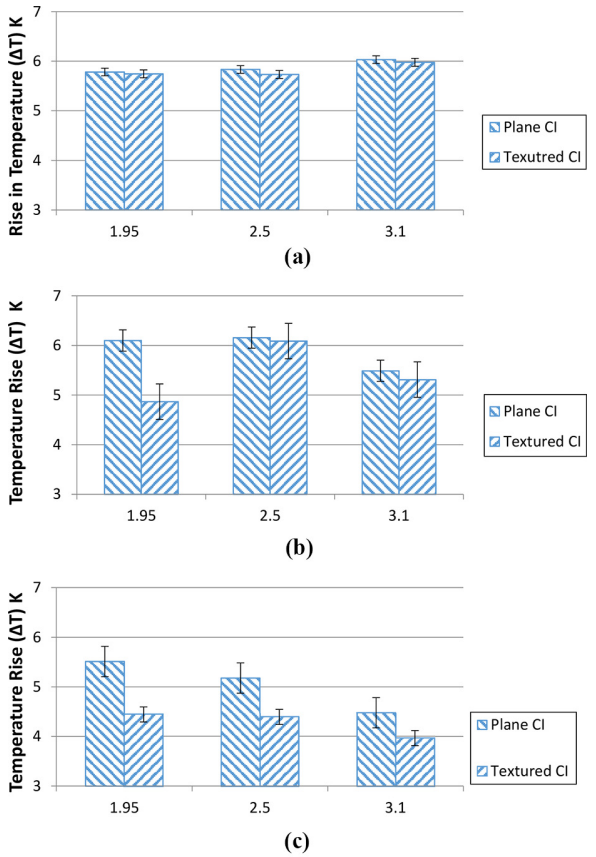
- There was a significant improvement in friction coefficient at the interface formed by textured surface CI disc and pin of

Figure 11 Sound (dB) vs Pressure (MPa)



Notes: (a) Pressure (MPa) at speed constant: 2m/sec; (b) pressure (MPa) at speed constant: 6m/sec; (c) pressure (MPa) at speed constant: 10 m/sec

Figure 12 Temperature rise (ΔT) K vs pressure (MPa)



Notes: (a) Pressure (MPa) at speed constant: 2m/sec; (b) pressure (MPa) at speed constant: 6m/sec; (c) pressure (MPa) at speed constant: 10 m/sec

chrome steel, under the fully flooded lubrication condition. The lubricating oil trapped in surface textures helped to reduce shear force, which in turn reduced friction coefficient. Also, the friction coefficient continuously decreased with the increase in the contact pressure up to a certain limit and then remained steady at a minimum value of 0.021.

- Further, the textured surface CI disc showed less wear. It was because of the lubricating oil trapped in the textures which helped in reducing wear. These textures on the surface act as a cushion. Also, the presence of the oil pockets due to the textured surface of the disc reduces the contact area of the tribopair, which in turn reduces the metal-to-metal contact area. The overall effect was to reduce wear significantly in case of tribopair formed by textured surface CI disc with steel pin (chrome steel).
- The minimum and maximum vibration amplitudes observed in case of tribopair formed by textured surface CI disc with a pin (chrome steel) were 38 and 48 μm, respectively. Plain surface CI disc has greater vibration amplitude in comparison to textured surface CI disc. It was because of the thin layer lubrication and more metal-to-metal contact.
- The temperature rise at the contact area of the tribopair formed by (i) plain surface CI disc, (ii) textured surface CI disc, with steel pin, was observed from 4 K to 6 K. Negligible temperature rise was observed in case of the textured surface CI disc as compared to the plain surface CI disc and steel pin. Overall it was observed that there was a negligible rise in temperature due to fully flooded lubrication conditions surrounding the interface.
- The sound values recorded during the experiments were between 66 and 73 dB. Plain CI disc generated more noise as compared textured surface CI disc. Presence of oil in tiny pockets (surface textures) and continuous flow of lubricating oil acted as a cushion. Hence, there was less metal-to-metal contact, which helped in less generation of sound in case of the textured surface CI disc.

Table I Summary results of disc

Element	Weight %	Atomic %
Carbon	17.66	49.35
Silicon	1.93	2.30
Iron	80.42	48.35

Table II Results summary: chrome steel pin

Element	Wt.%	Wt.% σ	Atomic percentage
C	20.110	0.575	42.636
O	17.244	0.543	27.446
Si	2.597	0.104	2.354
Cr	5.381	0.182	2.635
Fe	54.668	0.585	24.928

References

Amini, S., Hosseinabadi, N.H. and Sajjady, S.A. (2016), "Experimental study on effect of micro textured surfaces generated by ultrasonic vibration assisted face turning on friction and wear performance", *Applied Surface Science*, Vol. 390, pp. 633-648.

Bijwe, J., Gupta, M.K., Parida, T. and Trivedi, P. (2015), "Design and development of advanced polymer composites as high performance tribo-materials based on blends of PEK and ABPBI", *Wear*, Vols 342/343, pp. 65-76.

Cho, M. (2016), "Friction and wear of a hybrid surface texturing of polyphenylene sulfide-filled micro pores", *Wear*, Vols 346/347, pp. 158-167.

da Silva, L.R.R. and Costa, H.L. (2017), "Tribological behavior of gray cast iron textured by maskless electrochemical texturing", *Wear*, Vols 376/377, pp. 1601-1610.

Djoufack, M.H., Maya, U., Repphun, G., Brögelmann, T. and Bobzin, K. (2015), "Wear behaviour of hydrogenated DLC in a pin-on-disc model test under lubrication with different diesel fuel types", *Tribology International*, Vol. 92, pp. 12-20.

Dunn, A., Wlodarczyk, K.L., Carstensen, J.V., Hansen, E.B., Gabzdyl, J., Harrison, P.M., Shephard, J.D. and Hand, D.P. (2015), "Laser surface texturing for high friction contacts", *Applied Surface Science*, Vol. 357, pp. 2313-2319.

Etsion, I. and Sher, E. (2009), "Improving fuel efficiency with laser surface textured piston rings", *Tribology International*, Vol. 42 No. 4, pp. 542-547.

Gachot, C., Rosenkranz, A., Hsu, S.M. and Costa, H.L. (2017), "A critical assessment of surface texturing for friction and wear improvement", *Wear*, Vols 372/373, pp. 21-41.

Grabon, W., Pawlus, P., Wos, S., Koszela, W. and Wieczorowski, M. (2017), "Effects of honed cylinder liner surface texture on tribological properties of piston ring-liner assembly in short time tests", *Tribology International*, Vol. 113, pp. 137-148.

Hu, J. and Xu, H. (2016), "Friction and wear behavior analysis of the stainless steel surface fabricated by laser texturing underwater", *Tribology International*, Vol. 102, pp. 371-377.

Jones, K. and Schmid, S.R. (2016), "Experimental investigation of laser texturing and its effect on friction and lubrication", *Procedia Manufacturing*, Vol. 5, pp. 568-577.

Kim, B., Chae, Y.H. and Choi, H.S. (2014), "Effects of surface texturing on the frictional behavior of cast iron surfaces", *Tribology International*, Vol. 70, pp. 128-135.

Lal, R. and Singh, R.C. (2018), "Experimental comparative study of chrome steel pin with and without chrome plated cast iron disc in situ fully flooded interface lubrication", *Surface Topography, Metrology and Properties*, Vol. 6.

Li, K., Yao, Z., Hu, Y. and Gu, W. (2014), "Friction and wear performance of laser peen textured surface under starved lubrication", *Tribology International*, Vol. 77, pp. 97-105.

Luo, K.Y., Wang, C.Y., Li, Y.M., Luo, M., Huang, S., Hua, X.J. and Lu, J.Z. (2014), "Effects of laser shock peening and groove spacing on the wear behavior of non-smooth surface fabricated by laser surface texturing", *Applied Surface Science*, Vol. 313, pp. 600-606.

Meng, R., Deng, J., Liu, Y., Duan, R. and Zhang, G. (2018), "Improving tribological performance of cemented carbides by combining laser surface texturing and W-S-C solid lubricant coating", *International Journal of Refractory Metals & Hard Materials*, Vol. 72, pp. 163-171.

Mo, J.L., Wanga, Z.G., Chen, G.X., Shao, T.M., Zhu, M.H. and Zhou, Z.R. (2013), "The effect of groove-textured surface on friction and wear and friction-induced vibration and noise", *Wear*, Vol. 301 Nos 1/2, pp. 671-681.

Muthiah, S., Singh, R.C., Pathak, B.D. and Dhar, A. (2017), "Mechanical properties of thermoelectric n – type magnesium silicide synthesized employing in situ spark plasma reaction sintering", *Materials Research Express*, Vol. 4 No. 7, pp. 075507

Olofinjana, B., Lorenzo-Martin, C., Ajayi, O.O. and Ajayi, E.O. (2015), "Effect of laser surface texturing (LST) on tribochemical films dynamics and friction and wear performance", *Wear*, Vols 332/333, pp. 1225-1230.

Pandey, R.K., Tandon, N. and Singh, A.K. (2013), "Fuel saving in IC engine by surface texturing of piston rings", National Conference on Recent Advancements in Mechanical Engineering (NCRAME-2013).

Qu, Z., Wei, K., He, Q., He, R., Pei, Y., Wang, S. and Fang, D. (2018), "High temperature fracture toughness and residual stress in thermal barrier coatings evaluated by an in-situ indentation method", *Ceramics International*, Vol. 44 No. 7, pp. 7926-7929.

Rana, R., Walia, R.S. and Lata, S. (2018), "Development and investigation of hybrid electric discharge machining electrode process", *Materials Today Proceedings*, Vol. 5 No. 2, pp. 3936-3942.

Ryk, G. and Etsion, I. (2006), "Testing piston rings with partial laser surface texturing for friction reduction", *Wear*, Vol. 261 Nos 7/8, pp. 792-796.

Sharma, K.V., Singh, R.C. and Chaudhary, R. (2017), "Effect of flyash particles with aluminium melt on the wear of aluminium metal matrix composites", *Engineering Science and Technology, an International Journal*, Vol. 20 No. 4, pp. 1318 -1323.

Singh, R.C., Pandey, R.K., Rooplal, R. and MS and Maji, S. (2016), "Tribological performance analysis of textured steel surfaces under lubricating conditions", *Surface Topography, Metrology and Properties*, Vol. 4.

Sudeep, U., Pandey, R.K. and Tandon, N. (2013), "Effects of surface texturing on friction and vibration behaviors of sliding lubricated concentrated point contacts under linear reciprocating motion", *Tribology International*, Vol. 62, pp. 198-207.

Taie, I., Al-Shahrani, A., Qari, N., Fihri, A., Al-Obaid, W. and Alabedi, G. (2018), "High temperature corrosion resistant coatings for gas flare systems", *Ceramics International*, Vol. 44 No. 5, pp. 5124-5130.

Van Kuilenburg, J., Masen, M.A., Groenendijk, M.N.W., Bana, V., and Van der Heide, E. (2012), "An experimental study on the relation between surface texture and tactile friction", *Tribology International*, Vol. 48, pp. 15-21.

Vladescu, S.-C., Olver, A.V., Pegg, I.G. and Reddyhoff, T. (2016), "Combined friction and wear reduction in a reciprocating contact through laser surface texturing", *Wear*, Vol. 358-359, pp. 51-61.

Wos, S., Koszela, W. and Pawlus, P. (2017), "The effect of both surfaces textured on improvement of tribological properties of sliding elements", *Tribology International*, Vol. 113, pp. 182-188.

Wos, S., Koszela, W., Pawlus, P., Drabik, J. and Rogos, E. (2018), "Effects of surface texturing and kind of lubricant on the coefficient of friction at ambient and elevated temperatures", *Tribology International*, Vol. 117, pp. 174-179.

Xing, Y., Deng, J., Wua, Z. and Wu, F. (2017), "High friction and low wear properties of laser-textured ceramic surface under dry friction", *Optics and Laser Technology*, Vol. 93, pp. 24-32.

Yamaguchi, K., Takada, Y., Tsukuda, Y., Ota, M., Egashira, K. and Morita, T. (2016), "Friction characteristics of textured surface created by electrical discharge machining under lubrication", *Procedia CIRP*, Vol. 42, pp. 662 -667.

Zenebe Segu, D. and Hwang, P. (2015), "Friction control by multi-shape textured surface under pin-on-disc test", *Tribology International*, Vol. 91, pp. 111-117.

Further reading

Tang, W., Zhou, Y., Zhu, H. and Yang, H. (2013), "The effect of surface texturing on reducing the friction and wear of steel under lubricated sliding contact", *Applied Surface Science*, Vol. 273, pp. 199-204.

Corresponding author

Roop Lal can be contacted at: rooplalrana@gmail.com

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**

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/TCSII.2019.2925148, IEEE Transactions on Circuits and Systems II: Express Briefs

Isolation Improvement of MIMO Antenna using a Novel Flower Shaped Metamaterial Absorber at 5.5 GHz WiMAX band

Priyanka Garg and Priyanka Jain

***Abstract*—This article demonstrates the use of metamaterial absorber (MA) to achieve high isolation between two patch antennas in a 2-element MIMO system operating at 5.5 GHz resonant frequency useful for WiMAX application. The proposed flower shaped MA, designed on a $9\times9mm^2$ FR-4 substrate with 1 mm thickness, exhibits near unity normalized impedance at 5.5 GHz with an absorptivity of 98.7 %. A 4 element array of the MA is arranged in the form of a line in the middle of the two radiating patches in order to suppress the propagation of surface current between them at the operating frequency. Using the proposed flower shaped MA, an isolation of nearly 35 dB is achieved. The MIMO structure is studied in terms of return loss, isolation, overall gain, radiation pattern, Envelope Correlation Coefficient (ECC), Diversity Gain (DG), and Total Active Reflection Coefficient (TARC) etc. The structure is finally fabricated and measured to show good agreement with the simulated results.**

***Keywords*—Metamaterial absorber, MIMO antenna, WiMAX, isolation, mutual coupling reduction**

I. Introduction

Use of MIMO (multiple-input-multiple-output) antennas has increased in modern wireless communication systems as they offer significant improvement in link range as well as data throughput within the available bandwidth and power [1], [2]. Due to increased demand of miniaturization, mutual coupling reduction in MIMO antenna systems is the major issue of concern. Various techniques are available in literature to overcome this problem [3], [4]. In [3], a T-shaped stub is introduced in the ground plane to improve the isolation in UWB MIMO antenna system, whereas a carbon film is proposed in [4] which could absorb the interference between the antennas.

Metamaterial are the artificial structures that are being extensively used to improve the characteristics of various microwave components [5]. In this paper, the use of metamaterial based absorber for mutual coupling reduction in MIMO antenna has been explored. Metamaterial absorber was introduced by Landy *et al.* [6] in 2008. Such devices can start to resonate over a certain frequency range when an electromagnetic waves is incident upon them achieving near unity or so-called "perfect" absorption [7]. Such metamaterial absorbers are being utilized for several applications such as radar cross section (RCS) reduction [8], energy harvesting [9], sensing [10] and so on.

Priyanka Garg and Priyanka Jain are with the Department of Electronics & Communication Engineering, Delhi Technological University, New Delhi-110042, INDIA, e-mail: priyankagarg_phd2k16@dtu.ac.in, priyajain2000@rediffmail.com

TABLE I. A comparison of proposed technique with previously presented ones

Reference	Technique	Isolation (dB)	Frequency [GHz]
[14]	Decoupling Structure	35	2.67
[15]	Patch separation	33	6.8
[16]	Modified Serpentine	34	2.45
[17]	Decoupling Structure	30	2.5
[11]	Metamaterial Absorber	34	5.29
[12]	Metamaterial Absorber	30	5.1
Proposed	**Metamaterial Absorber**	**43.71**	**5.5**

However, using metamaterial absorber for isolation improvement in MIMO antennas has not been explored much [11]–[13]. A comparison of proposed technique with previously presented ones is given in Table I, taking into account the structures operating at single resonant frequency.

In this paper, a patch antenna is designed to operate at 5.5 GHz WiMAX frequency which is arranged in a 2 element MIMO array and its characteristics are studied. Further, a flower-shaped metamaterial absorber is designed and analysed using periodic boundaries and floquet ports. The proposed MA exhibits 98.7 % absorptivity at 5.5 GHz. An array of 4 elements of the designed absorber is then placed in the middle of the MIMO antenna system and its characteristics are studied. The use of MA helped in achieving a simulated isolation level of more than 30 dB. The structure is further analysed to examined its applicability as a MIMO antenna in terms of ECC, DG and TARC. The design is finally fabricated and measured to achieve an isolation level of more than 40 dB. All the designing and simulation is done using Ansys High freqyency structure simulator (HFSS) [18].

II. Proposed metamaterial absorber design and simulation

A metamaterial based absorber is designed on a 1 mm thick FR-4 substrate with dielectric constant(ϵ_r) of 4.4 and loss tangent(δ) of 0.02 with a unit cell size of $9\times9\ mm^2$ which is much less than $\lambda/4$ where λ is the wavelength at the operating frequency, making the structure effectively homogeneous. The proposed MA unit cell consists of four square rings split at the corner, arranged in a form of a symmetrical flower shape, as shown in fig. 1(a). The evolution of the proposed MA, shown in fig. 2, starts with a conventional square ring operating around 9 GHz (fig. 2(I)). Then a split was introduced at the corner of the ring, shown in fig. 2(II), to increase the equivalent capacitance due to which the operating frequency shifted towards lower side, however, the absorptivity obtained was less than 50%.

1549-7747 (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/TCSII.2019.2925148, IEEE Transactions on Circuits and Systems II: Express Briefs

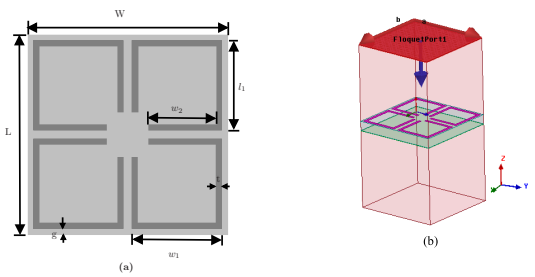


Fig. 1. (a) Schematic of the proposed metamaterial absorber unit cell. (b) Simulation set up of the proposed MA unit cell

Further, to obtain four fold symmetry, a flower shaped design was constructed by rotating the split square ring 90° in the clockwise direction. It resulted in a significant improvement in the absorptivity due to increased electric and magnetic coupling between the cells. Finally, the design was scaled down and optimized to provide an absorptivity of 98.7% at 5.5 GHz (fig. 2(III)).

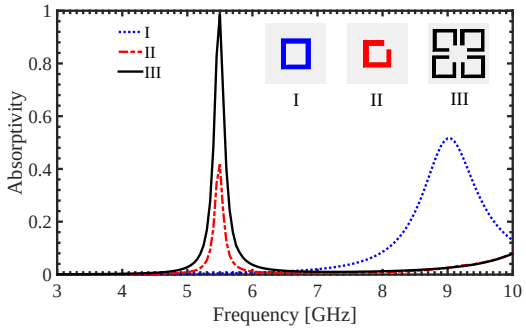


Fig. 2. Absorptivities of different metamaterial absorber shapes.

The design parameters of MA as indicated in fig. 1(a) are given as W=9mm, L=9mm, l_1=4.05mm, w_1=4.05mm, w_2=3.2mm, g=0.2mm, t=0.3mm. The MA is simulated by applying floquet port and periodic boundary conditions in Ansys HFSS, shown in fig. 1(b). The absorptivity of the proposed MA is obtained using Eq. 1

$$A(\omega) = 1 - |S_{11}(\omega)|^2 - |S_{21}(\omega)|^2 \tag{1}$$

Where, $|S_{11}(\omega)|^2$ and $|S_{21}(\omega)|^2$ represents the reflectivity and transitivity respectively with respect to frequency ω. Since, the bottom of the substrate is completely laminated by copper so all the power is reflected back from the bottom which indicated zero transmission, thus, $|S_{21}(\omega)|^2$=0. Therefore, Eq. 1 reduces to

$$A(\omega) = 1 - |S_{11}(\omega)|^2 \tag{2}$$

Thus, to obtain near unity absorption, Reflected power from the unit cell has to be minimized. The reflectivity from a medium for a normally incident electromagnetic wave is given by eq. 3 [19].

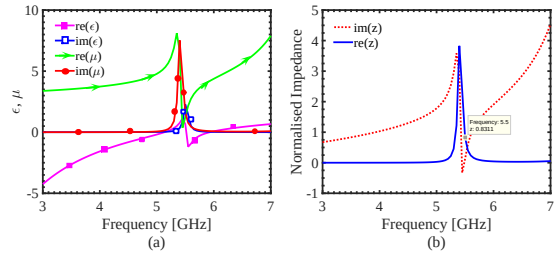


Fig. 3. (a) Material parameters of metamaterial absorber (b) Normalized impedance of the metamaterial absorber

$$R = \left|\frac{\mu_r - n}{\mu_r + n}\right|^2 \tag{3}$$

where, μ_r and ϵ_r are relative permeability and permittivity of the medium and n is the refractive index of medium given by $\sqrt{\mu_r \epsilon_r}$. When ϵ_r and μ_r are equal, reflectivity (R) is ideally zero. Fig. 3(a) depicts the real and imaginary part of permittivity and permeability. At 5.5 GHz, the value of permittivity is 0.6612+j2.269 and permeability is 1.24+j1.604 which are almost equal in magnitude, thus resulting in very low value of reflectivity.

Now, the normalized impedance (normalized to 377 ohms i.e. the free space impedance) of the absorber is given by Eq. 4

$$Z_{nor} = \sqrt{(\mu_r/\epsilon_r)} \tag{4}$$

If both the ϵ_r and μ_r are equal, then the normalized impedance of the absorber Z_{nor} is ideally unity which implies perfect impedance matching of the structure with the free space, necessary to obtain perfect absorption. Fig. 3(b) shows the real and imaginary part of normalized impedance. The real part of normalised impedance is observed to be 0.8311 at the operating frequency 5.5 GHz which is close to the free space impedance and its imaginary part is close to zero, thus obtaining near perfect absorption.

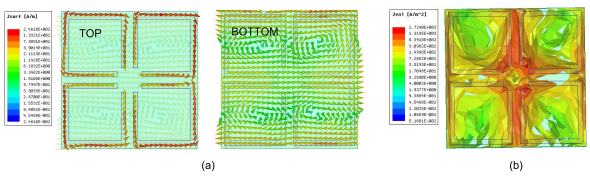


Fig. 4. (a) Surface current density plot of MA at 5.5 GHz (Top view and Bottom view) (b) Volume current density plot of MA at 5.5 GHz.

The absorption phenomenon is further explained using the surface current density and volume current density plot of the MA at 5.5 GHz, as shown in fig. 4(a) and (b) respectively. Fig. 4(a) shows that at the resonant frequency, current flowing on top and bottom surfaces are anti-parallel to each other forming current loops in the direction perpendicular to the applied magnetic field. This indicates magnetic excitation in

1549-7747 (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/TCSII.2019.2925148, IEEE Transactions on Circuits and Systems II: Express Briefs

the structure which manipulates magnetic permeability of the structure. Fig. 4(b) shows volume current distribution of the MA at the resonant frequency where near unity absorption takes place. This indicates the occurrence of dielectric losses within the FR-4 substrate. Therefore, the fig. 4 clearly shows that electric and magnetic excitation leads to strong electro-magnetic absorption at 5.5 GHz.

The proposed flower shaped absorber is engineered as an effective metamaterial medium (fulfilling the inclusion size requirement of an effectively homogeneous structure) to provide nearly equal value of relative permittivity and permeability (not found in any naturally occurring material) by properly optimizing the size and dimensions of the structure; therefore, it can be called as a true metamaterial.

III. MIMO ANTENNA DESIGN AND SIMULATION

First of all a conventional microstrip patch antenna with patch dimensions $12.2\times16.1mm^2$ is designed on a 1 mm thick FR-4 substrate using a quarter wave transformer feed to operate at 5.5 GHz resonant frequency. The antenna is further arranged in a 1×2 array system. The centre to centre spacing between the two antennas is kept 30 mm which is nearly equivalent to $\lambda_0/2$, where λ_0 is the wavelength at operating frequency. To improve the isolation between the two patch antennas, a linear array of four elements of the proposed MA is placed in the middle of the two patch antennas operating at 5.5 GHz, shown as an inset in the fig. 5. Such a structure can be termed as a "metamaterial-inspired" kind, since its properties are determined by the interaction with only a 4-element array of the proposed absorber, rather than with an effective metamaterial medium.

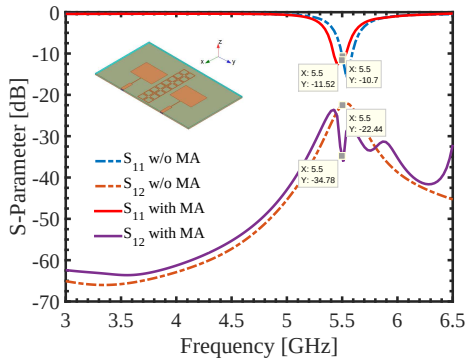


Fig. 5. Simulated S-parameters with and without the proposed MA along with 3D view of the proposed MIMO antenna with flower shaped MA.

The proposed MIMO structure is simulated using Ansys HFSS and the results are analysed. Fig. 5 shows the simulated S-parameters of the MIMO antenna with and without the MA. Due to symmetry of the design, S_{11} is equivalent to S_{22} and S_{12} is equivalent to S_{21}, therefore, only S_{11} and S_{12} are displayed in the Fig. 5. Results show that at 5.5 GHz, the reflection coefficient (S_{11}) remains almost same for both the cases whereas the parameter S_{12} decreases from -22.46 dB to -34.87 dB showing a mutual coupling reduction by 12.41 dB as compared to the conventional MIMO antenna system. Thus a very high amount of isolation can be achieved with the help of proposed metamaterial absorber.

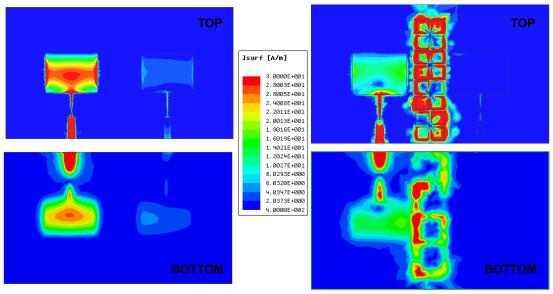


Fig. 6. Surface current distribution at 5.5 GHz (a) without MA, (b) with MA.

The operating mechanism for mutual coupling reduction using MA can be further understood using surface current distribution plot. Fig. 6(a) and (b) respectively shows the surface current distribution of the MIMO antenna without and with metamaterial absorber at 5.5 GHz obtained by exciting antenna 1 and terminating antenna 2 by matched load. It can be seen that without the use of MA, some amount of the surface current gets coupled from antenna 1 to antenna 2. Whereas, in case of MIMO antenna with MA, the propagating current on the top surface and bottom ground plane excites the absorber array in such a way that an anti-parallel current starts propagating between the top flower shape structure and the ground beneath it. This phenomenon results in trapping of the magnetic field within the MA array, thus not allowing the surface current to flow further, as shown in fig. 6(b). Therefore, use of the MA stops the propagation of electromagnetic energy from antenna 1 to antenna 2, thus resulting in high level of isolation between them.

Furthermore, the simulated 2D radiation patterns of MIMO antenna are studied at 5.5 GHz to observe the effect of MA on the radiation performance of MIMO structure. Fig. 7(a),(b) and (c) shows the 2D far field radiation pattern of proposed structure on xz, yz and xy planes respectively. From the fig. 7 it can be concluded that there is no significant effect on the radiation-pattern of the antenna after placing the MA, except that the magnitude of the main lobe is increased a little. This shows an improvement in the antenna parameters in terms of radiation efficiency and gain along with preserving the radiation characteristics as an effect of the reduced mutual coupling between the antennas. Fig. 8 shows the peak gain and radiation efficiency plot of MIMO structure. The graph shows a peak gain of 7.53 dB without MA and 7.74 dB with MA at 5.5 GHz. Also, there is an increase in radiation efficiency from 66.64 % to 68.03%.

IV. MIMO DIVERSITY PERFORMANCE

The diversity performance of MIMO antenna is analysed on the basis of different parameters such as Envelope Correlation Coefficient (ECC), Diversity Gain (DG), and Total Active

1549-7747 (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/TCSII.2019.2925148, IEEE Transactions on Circuits and Systems II: Express Briefs

$$ECC = \left| \frac{|S_{11}^* S_{12} + S_{12}^* S_{22}|}{|(1-|S_{11}|^2-|S_{21}|^2)(1-|S_{12}|^2-|S_{22}|^2)(\eta_{rad,1}\eta_{rad,2})|^{1/2}} \right|^2 \quad (5)$$

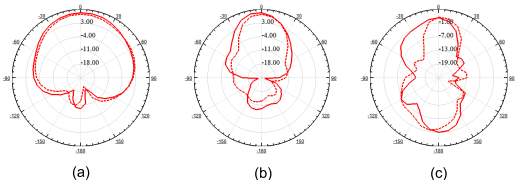


Fig. 7. MIMO antenna far field radiation pattern at 5.5 GHz with (-solid line) and without (–dashed line) MA on (a) xz-plane (b) yz-plane (c) xy-plane.

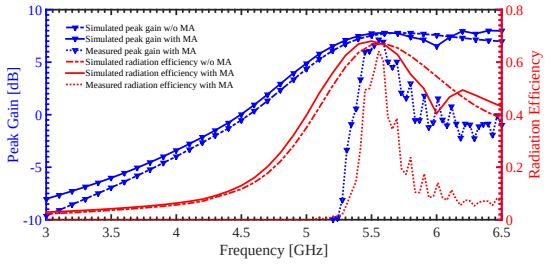


Fig. 8. Peak gain and radiation efficiency plot of MIMO with and without MA

Reflection Co-efficient (TARC) [20]. For a practical MIMO antenna, it is required to have minimum signal correlation between the antenna elements which can be analysed using ECC. It can be obtained using the S-parameters and radiation efficiency, for a 2-port MIMO system, given by eq. 5 [21], where, $\eta_{rad,1}$ and $\eta_{rad,2}$ are the radiation efficiencies of antenna 1 and antenna 2. For a satisfactory performance

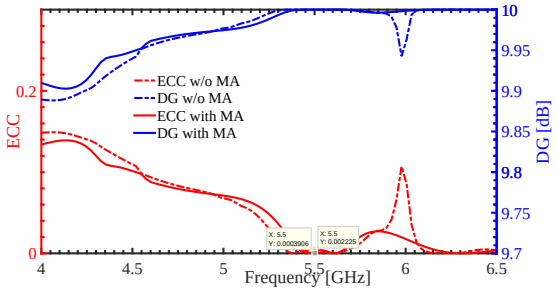


Fig. 9. Simulated ECC and DG of MIMO antenna with MA.

MIMO antenna should have ECC less than 0.5. Fig. 9 shows a simulated ECC of less than 0.05 for the operating band which shows low correlation between the two antennas. The diversity gain can be obtained using eq. 6 which is also shown in fig. 9. The value of diversity gain is observed to be close to 10 dB, which indicates good diversity performance.

$$DG = 10 \times \sqrt{(1 - ECC^2)} \quad (6)$$

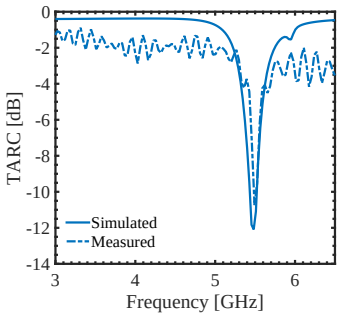


Fig. 10. Simulated and measured TARC of MIMO antenna with MA.

Total Active Reflection Co-efficient (TARC) is another parameters to define the diversity performance of MIMO structure. It is the square root of the ratio of total reflected power to the total incident power and hence, it shows the apparent return loss of the overall MIMO antenna system. For a side-by-side 2-port MIMO system, it is given by eq. 7.

$$TARC = \sqrt{\frac{(S_{11}+S_{12})^2 + (S_{22}+S_{21})^2}{2}} \quad (7)$$

It is expected to have a TARC value of less than 0 dB for a MIMO system to work properly. Fig. 10 shows the plot of TARC with respect to frequency and it is observed to have a TARC value of less than -10 dB for the desired frequency which is well below the expected value.

V. Experimental Results and Discussion

After studying all the simulated results, antenna is finally fabricated and tested using Keysight N9914A vector network analyzer. The fabricated prototype of the proposed structure along with the measured S-parameters compared to the simulated ones as a function of frequency are shown in fig. 11 Result shows that a measured isolation level of -43.71 dB is achieved between the two conventional patch antennas at 5.5 GHz WiMAX frequency by using an array of proposed metamaterial absorber between them. The values of measured gain and radiation efficiency, shown in the fig. 8, are observed to be 6.28 dB and 56% respectively at the resonant frequency 5.5 GHz. Measured value of ECC calculated at 5.5 GHz is 0.00043. Measured TARC is also indicated in fig. 10.

Table II gives a performance comparison of the MIMO antenna with and without the proposed MA at 5.5 GHz.

VI. Conclusion

The effect of utilizing a metamaterial based absorber for mutual coupling reduction in a MIMO antenna system is examined. A novel flower shaped metamaterial absorber is designed with an absorptivity of 98.7 % and placed between

1549-7747 (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/TCSII.2019.2925148, IEEE Transactions on Circuits and Systems II: Express Briefs

TABLE II. PERFORMANCE COMPARISON OF THE MIMO ANTENNA WITH AND WITHOUT THE PROPOSED MA AT 5.5 GHZ

MIMO antenna	Simulated S_{11} [dB]	Simulated S_{12} [dB]	Measured S_{11} [dB]	Measured S_{12} [dB]	Peak Gain [dB]	Radiation Efficiency
Without MA	-10.68	-22.46	-11.2	-22.1	7.53	66.64 %
With MA	-11.5	-34.77	-10.40	-43.71	7.74	68.03 %

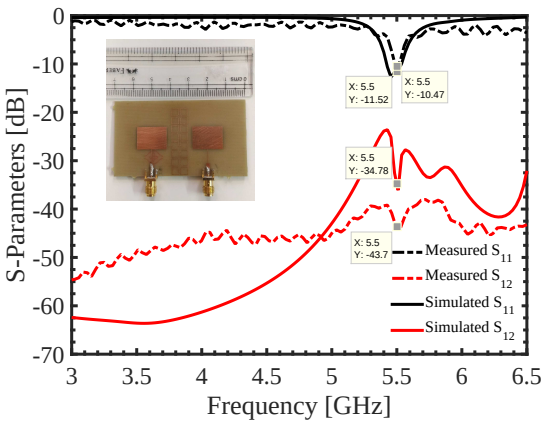


Fig. 11. Measured and simulated S-parameters with fabricated prototype of proposed structure.

the two conventional patch antennas to observe a simulated isolation level of 34.77 dB and measured isolation of 43.71 dB at the 5.5 GHz WiMAX frequency without effecting the return loss. An improvement in peak gain from 7.53 dB to 7.74 dB and radiation efficiency from 66.64 % to 68.03% is also observed. Other parameters like ECC, diversity gain and TARC are also studied and are found to be fairly below the expected limit of wireless communication application. Therefore, the proposed structure is quite suitable for MIMO application showing high level of isolation and good amount of radiation efficiency and peak gain.

VII. ACKNOWLEDGEMENT

Authors would like to thank Prof. Vibhakar Shrimali and Mr. Padam Saini of G. B. Pant Govt. Engineering College, New Delhi, for their support in providing the measurement facilities.

REFERENCES

[1] Da-Shan Shiu, G. J. Foschini, M. J. Gans, and J. M. Kahn, "Fading correlation and its effect on the capacity of multielement antenna systems," *IEEE Transactions on Communications*, vol. 48, no. 3, pp. 502–513, March 2000.

[2] A. F. Molisch, M. Z. Win, Yang-Seok Choi, and J. H. Winters, "Capacity of MIMO systems with antenna selection," *IEEE Transactions on Wireless Communications*, vol. 4, no. 4, pp. 1759–1772, July 2005.

[3] R. Chandel and A. K. Gautam, "Compact MIMO/diversity slot antenna for UWB applications with band-notched characteristic," *Electronics Letters*, vol. 52, no. 5, pp. 336–338, 2016.

[4] G. Lin, C. Sung, J. Chen, L. Chen, and M. Houng, "Isolation improvement in UWB MIMO antenna system using carbon black film," *IEEE Antennas and Wireless Propagation Letters*, vol. 16, pp. 222–225, 2017.

[5] I. Buriak, V. Zhurba, G. Vorobjov, V. Kulizhko, O. Kononov, and O. Rybalko, "Metamaterials: Theory, classification and application strategies," *Journal Of Nano- And Electronic Physics*, vol. 8, no. 4 (2), pp. 04 088–1, 2016.

[6] N. I. Landy, S. Sajuyigbe, J. J. Mock, D. R. Smith, and W. J. Padilla, "Perfect metamaterial absorber," *Physical Review Letters*, vol. 100, no. 4, pp. 207–402, May 2008.

[7] S. Ji, C. Jiang, J. Zhao, X. Zhang, and Q. He, "Design of a polarization-insensitive triple-band metamaterial absorber," *Optics Communications*, vol. 432, pp. 65 – 70, 2019.

[8] S. Narayan, B. Sangeetha, T. V. Sruthi, V. Shambulingappa, and R. U. Nair, "Design of low observable antenna using active hybrid-element FSS structure for stealth applications," *AEU - International Journal of Electronics and Communications*, vol. 80, pp. 137 – 143, 2017.

[9] M. Bagmanc, M. Karaaslan, E. Unal, O. Akgol, F. Karadag, and C. Sabah, "Broad-band polarization-independent metamaterial absorber for solar energy harvesting applications," *Physica E: Low-dimensional Systems and Nanostructures*, vol. 90, pp. 1 – 6, 2017.

[10] M. Bakr, M. Karaaslan, E. Unal, O. Akgol, and C. Sabah, "Microwave metamaterial absorber for sensing applications," *Opto-Electronics Review*, vol. 25, no. 4, pp. 318 – 325, 2017.

[11] M. Agarwal and M. K. Meshram, "Isolation improvement of 5 GHz WLAN antenna array using metamaterial absorber," in *2016 URSI Asia-Pacific Radio Science Conference (URSI AP-RASC)*, Aug 2016, pp. 1050–1053.

[12] Q. Zhang, Y. Jin, J. Feng, X. Lv, and L. Si, "Mutual coupling reduction of microstrip antenna array using metamaterial absorber," in *2015 IEEE MTT-S International Microwave Workshop Series on Advanced Materials and Processes for RF and THz Applications (IMWS-AMP)*, July 2015, pp. 1–3.

[13] H. S. Lee and H. M. Lee, "Isolation improvement between loop antennas with absorber cells," in *2011 IEEE International Symposium on Antennas and Propagation (APSURSI)*, July 2011, pp. 1735–1738.

[14] S. R. Thummaluru and R. K. Chaudhary, "Mu-negative metamaterial filter-based isolation technique for MIMO antennas," *Electronics Letters*, vol. 53, no. 10, pp. 644–646, 2017.

[15] K. J. Babu, R. W. Aldhaheri, M. Talha, and I. SAlruhaili, "Design of a compact two element MIMO antenna system with improved isolation," *Progress In Electromagnetics Research*, vol. 48, pp. 27–32, 2014.

[16] H. Arun, A. K. Sarma, M. Kanagasabai, S. Velan, C. Raviteja, and M. G. N. Alsath, "Deployment of modified serpentine structure for mutual coupling reduction in MIMO antennas," *IEEE Antennas and Wireless Propagation Letters*, vol. 13, pp. 277–280, 2014.

[17] C. Xue, X. Y. Zhang, Y. F. Cao, Z. Hou, and C. F. Ding, "MIMO antenna using hybrid electric and magnetic coupling for isolation enhancement," *IEEE Transactions on Antennas and Propagation*, vol. 65, no. 10, pp. 5162–5170, Oct 2017.

[18] Ansoft, "HFSS simulator version 14.0."

[19] C. M. Watts, X. Liu, and W. J. Padilla, "Metamaterial electromagnetic wave absorbers," *Advanced materials*, vol. 24, no. 23, pp. OP98–OP120, 2012.

[20] R. Chandel, A. K. Gautam, and K. Rambabu, "Tapered fed compact UWB MIMO-diversity antenna with dual band-notched characteristics," *IEEE Transactions on Antennas and Propagation*, vol. 66, no. 4, pp. 1677–1684, April 2018.

[21] G. Srivastava and A. Mohan, "Compact MIMO slot antenna for UWB applications," *IEEE Antennas and Wireless Propagation Letters*, vol. 15, pp. 1057–1060, 2016.

1549-7747 (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.

Soft Computing
https://doi.org/10.1007/s00500-019-04169-y

METHODOLOGIES AND APPLICATION

Kernel intuitionistic fuzzy entropy clustering for MRI image segmentation

Dhirendra Kumar[1,2] · R. K. Agrawal[1] · Hanuman Verma[3]

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

Abstract
Fuzzy entropy clustering (FEC) is a variant of hard c-means clustering which utilizes the concept of entropy. However, the performance of the FEC method is sensitive to the noise and the fuzzy entropy parameter as it gives incorrect clustering and coincident cluster sometimes. In this work, a variant of the FEC method is proposed which incorporates advantage of intuitionistic fuzzy set and kernel distance measure termed as kernel intuitionistic fuzzy entropy c-means (KIFECM). While intuitionistic fuzzy set allows to handle uncertainty and vagueness associated with data, kernel distance measure helps to reveal the inherent nonlinear structures present in data without increasing the computational complexity. In this work, two popular intuitionistic fuzzy sets generators, Sugeno and Yager's negation function, have been utilized for generating intuitionistic fuzzy sets corresponding to data. The performance of the proposed method has been evaluated over two synthetic datasets, Iris dataset, publicly available simulated human brain MRI dataset and IBSR real human brain MRI dataset. The experimental results show the superior performance of the proposed KIFECM over FEC, FCM, IFCM, UPCA, PTFECM and KFEC in terms of several performance measures such as partition coefficient, partition entropy, average segmentation accuracy, dice score, Jaccard score, false positive ratio and false negative ratio.

Keywords Intuitionistic fuzzy sets · Fuzzy entropy clustering · Kernel distance measure · Image segmentation · Magnetic resonance imaging

1 Introduction

Clustering is an important unsupervised technique, which is utilized to partition observed unlabeled data based on intrinsic properties to reveal the structural information associated with it. There are plenty real-world applications such as document classification, content retrieval, image classification (Berry and Castellanos 2004; Jain et al. 1999; Rui et al. 1999), where clustering methods have been utilized. Hard c-means (HCM) clustering method, which is based on minimization of square error, is the most widely used conventional clustering method (Lloyd 1982). This clustering method assigns a data point to exactly one cluster. However, some of the real-world data may not have a good interpretation of the outcome of HCM method as a data point may belong to boundary of two or more clusters. Fuzzy c-means (FCM) (Bezdek 1981) address this problem by assigning a membership value in the interval [0 1] for a data point to each cluster. Clustering is one of the methods utilized to partition pixels of an image into a number of similar regions, which is also known as image segmentation. Image segmentation is quite often used by radiologist to analyze medical image data.

In advance medical era, there are many imaging modalities such as magnetic resonance imaging (MRI), computed tomography (CT), X-rays, mammogram, positron-emission tomography (PET), ultrasound, etc. which help to acquire the internal structure of human organs. For diagnosis of many diseases, human radiologists use their knowledge and expe-

Communicated by V. Loia.

✉ Dhirendra Kumar
dhirendrakumar@dtu.ac.in

R. K. Agrawal
rkajnu@gmail.com

Hanuman Verma
hv4231@gmail.com

[1] School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi, India

[2] Department of Applied Mathematics, Delhi Technological University, Delhi, India

[3] Department of Mathematics, Bareilly College, Bareilly, Uttar Pradesh, India

Published online: 27 June 2019

riences for correctly identifying the abnormalities with the help of these medical images. Nowadays, due to advancement and availability of imaging technologies, a large amount of imaging data are generated. The analysis of these imaging data requires more time due to limited availability of the experienced radiologists. The computer-aided diagnosis may help the radiologist to analyze more such images in less time.

MRI imaging is a popular modality for better understanding and analysis of a diseased organ and provides better resolution without bad effect on the tissue unlike X-rays and PET exposure, which harm the healthy tissues while scanning the body part for delineation. The main artifacts which affect the quality of the acquired MRI image during the acquisition are noise, intensity inhomogeneity and partial volume effect (Hoffman et al. 1979). Due to this artifact, the performance of segmentation of the MRI image may deteriorate. In the literature, many approaches have been proposed for MRI image segmentation such as level set (Alipour and Shanbehzadeh 2014; Li et al. 2011; Wang et al. 2010), graph cut (Chen et al. 2016), region growing (Sato et al. 2000), atlas-guided methods (Rohlfing et al. 2004), artificial neural network (ANN) (Reddick et al. 1997), Markov random field model (Held et al. 1997) and clustering (Balafar et al. 2010; Benaichouche et al. 2013; Cocosco et al. 1997; Huang et al. 2015; Ji et al. 2014; Kannan et al. 2013; Krinidis and Chatzis 2010; Kumar et al. 2018; Olabarriaga and Smeulders 2001; Pham et al. 2000; Qiu et al. 2013; Wang et al. 2013). Among these segmentation techniques, clustering based on fuzzy set theory (Zadeh 1965) known as fuzzy c-means (FCM) (Balafar 2014; Bezdek 1981) is found to be better in comparison with conventional hard clustering algorithm as it models characteristics related to MRI brain images. However, FCM clustering method is sensitive to noisy data. Many variants of FCM algorithm are proposed, which incorporate the spatial information while clustering (Nayak et al. 2015) to reduce the effect of noise.

In the literature, to handle noise, Krishnapuram and Keller proposed possibilistic c-means (PCM) algorithm (Krishnapuram and Keller 1993), which relaxes the constraints in FCM. However, the performance of segmentation depends on choice of parameters and sometimes it provides coincident clusters for datasets having a large number of data points. A new hybrid of FCM and PCM known as PFCM (Pal et al. 2005) is proposed to improve the performance of clustering with the assumptions that membership (relative typicality) and possibilistic (absolute typicality) are both important. Another FCM based clustering method integrated with the Shannon's entropy within the objective function is proposed in the research work (Kahali et al. 2018) to handle the class uncertainty.

In the FCM and its variants, the assignment of a data point to a given cluster is expressed in terms of a membership value in interval [0 1], which is a function of the distance between the data point and the cluster centroid. But, in case of real-world applications, It has been observed that the imprecise measurement and noise while acquiring the data point always arises an uncertainty with the localization of data points. Due to this, the computation of membership values of a data point to different clusters and centroids is imprecise (Pelekis et al. 2008). To overcome this problem, Atanassov suggested intuitionistic fuzzy sets (IFSs) theory (Atanassov 1986, 2003) which is an extension of fuzzy set theory. IFSs have capability to deal with uncertainty and vagueness associated with real data as it takes advantage of non-membership degree and hesitation degree along with membership degree for representing the real data (Szmidt and Kacprzyk 2000). Recently, clustering based on IFSs theory is utilized for image segmentation (Chaira 2011; Iakovidis et al. 2008; Kumar et al. 2018; Verma and Agrawal 2015; Verma et al. 2016; Xu and Wu 2010). Incorporation of intuitionistic fuzzy set theory for clustering process provides better performance which is more accurate, robust to noise and faster convergence in comparison with conventional FCM (Iakovidis et al. 2008).

Due to nonlinear structure present in real data, most of the existing clustering methods do not produce promising results. In the literature, there are research works which advocate the utilization of nonlinear transformation of original data to high-dimensional space to capture nonlinear structure present in real data. In this direction, many variants of FCM based on kernel-induced distance has been proposed in the literature (Chen and Zhang 2004; Graves and Pedrycz 2010; Li et al. 2018; Lin 2013) along with the incorporation of spatial information for image segmentation.

In general, soft clustering methods such as the FCM and its variants have higher partition entropy which decreases the confidence for taking decision (Schaaf and Kemp 1997; Verma et al. 2014; Yager 2000). To improve the performance of segmentation, fuzzy entropy clustering (FEC) (Tran and Wagner 2000) is suggested which is a variant of HCM with an additional term of fuzzy entropy. However, it is sensitive to both the noise and the fuzzy entropy parameter as it gives incorrect clustering and coincident clusters sometimes. To overcome this problem, a possibilistic type of fuzzy entropy c-means (PTFECM) algorithm, based on FEC and PCM, is suggested by Hai-Jun et al. (2011), which is similar to PFCM to relax the probabilistic constraints. The research work (Hai-Jun et al. 2011) overcomes the drawback of FEC methods but inherits the problem related to possibilistic clustering method. Another research work based on possibilistic theory, unsupervised possibilistic clustering approach (UPCA) (Yang and Wu 2006), is suggested to overcome the problem of FCM and PCM and proposed the cluster validity index coefficients such as the partition coefficient and partition entropy for possibilistic partition matrix. However, it produces coincident clusters for large datasets. Recently, a research work (Pham et al. 2018) proposed an improved particle swarm optimization (PSO)-based kernelized FEC method

for clustering with local spatial information and bias correction (KFECSB). Though, this method help in tuning the fuzzy entropy parameter but is time-consuming.

As discussed above, higher partition entropy will decrease the confidence while making decision based on the outcome of image segmentation. In order to increase the confidence of outcome of segmentation process, a variant of FEC method, which incorporates the advantages of IFSs and Gaussian kernel distance measure termed as intuitionistic fuzzy entropy c-means (KIFECM), is proposed in this paper. The objective function of the optimization problem of the proposed KIFECM method consists of three terms. The first term is the total clustering error in kernel space. The second term forces the membership to be maximum to increase the chances of getting non-trivial solution. The third term, fuzzy entropy factor, is included to obtain a good partition of IFSs. Due to the transformation of data points in higher-dimensional space in terms of kernel distance, the representational capability is increased which will reveal the inherent non-Euclidean structures in data without increasing the computational complexity. In this work, we have investigated Sugeno and Yager's negation function for generating the intuitionistic fuzzy set corresponding to real data.

Further, to check the efficacy of the proposed KIFECM method, experiments are performed on various real and synthetic datasets. The performance of the proposed method is compared quantitatively in terms of partition coefficient, partition entropy, average segmentation accuracy, dice score, Jaccard score, false positive ratio and false negative ratio with well-known existing methods such as FEC, FCM, IFCM, UPCA, PTFECM and a recent research work (Pham et al. 2018) referred as KFEC (Pham et al. 2018).

The remaining part of this paper is organized as follow. In Sect. 2, preliminaries and notations are included. The proposed method is described in Sect. 3. The experimental results on various synthetic and real data are presented in Sect. 4. Finally, Sect. 5 includes the conclusion.

2 Preliminaries and notations

The description of notations and related definitions used in this work are described here in this section.

2.1 Intuitionistic fuzzy set theory

Fuzzy set F defined over a set X can be represented as:

$$F = \{(x, \mu_F(x)) : x \in X\} \tag{1}$$

where $\mu_F(x)$ is membership function defined for an element $x \in X$ which represents the degree of belongingness (Zadeh 1965).

$\mu_F : X \rightarrow [0\ 1]$

Entropy of fuzzy set is analogous to entropy of probability distribution (Pal and Pal 1992). It is defined for a given fuzzy set F which contains members $x_j \in X$ with membership value $\mu_F(x_j)\ \forall j = 1, 2, \ldots, N$ as (Li and Mukaidono 1999):

$$E(F) = -\lambda \sum_{j=1}^{N} \mu_F(x_j) \log \mu_F(x_j) \tag{2}$$

where $\lambda > 0$ is fuzzy entropy factor. The fuzzy entropy $E(F)$ for a fuzzy set F represents the average value of uncertainty or average degree of non-membership (Li and Mukaidono 1999).

Intuitionistic fuzzy set P is an extension of fuzzy set over X is defined as (Atanassov 1986):

$$P = \{(x, \mu_P(x), \nu_P(x)) : x \in X \text{ and } 0 \leq \mu_P(x) + \nu_P(x) \leq 1\} \tag{3}$$

where $\mu_P(x)$ and $\nu_P(x)$ represent the degree of membership and non-membership values of an element x in the IFS P, respectively. There is an another uncertainty arises while defining the membership degree because of insufficient knowledge which is termed as hesitation. The hesitation degree $\pi_P(x)$ of an elements x in IFS P can be given as (Atanassov 1986):

$$\pi_P(x) = 1 - \mu_P(x) - \nu_P(x) \tag{4}$$

Usually, intuitionistic fuzzy generator is used to calculate the non-membership degrees for constructing the IFSs (Bustince et al. 2000).

A non-decreasing continuous function $g(\mu)$ can be termed as intuitionistic fuzzy generator if it follows:

$$g(\mu) \leq (1 - \mu) \text{ for all } \mu \in [0, 1] \text{ and } g(0) \leq 1 \text{ and } g(1) \leq 0$$

the general form of the negation function $NM(\mu)$ (non-membership function) for a given generating function $g(.)$ can be defined as (Bustince et al. 2000):

$$\text{NM}(\mu) = g^{-1}(g(1) - g(\mu)) \tag{5}$$

There are two famous generating function which are used to calculate the non-membership value for construction of the IFSs.

- Sugeno's negation function (SNF): A Sugeno class can be created with a given negation parameter β with help of the

following generating function (Murofushi and Sugeno 2000):

$$g(\mu) = \frac{1}{\beta}\log(1+\beta\mu) \tag{6}$$

using (5) and (6), the non-membership value $\nu_P(x)$ for a given membership value $\mu_P(x)$ can be calculated as follows:

$$\nu_P(x) = \mathrm{NM}(\mu_P(x)) = \frac{1-\mu_P(x)}{1+\beta\mu_P(x)},\ \beta > 0 \tag{7}$$

– Yager's negation function (YNF): A Yager class can be obtained with a given negation parameter β with help of the following generating function (Yager 1979, 1980):

$$g(x) = x^{\beta} \tag{8}$$

using (5) and (8), the non-membership value $\nu_P(x)$ for a given membership value $\mu_P(x)$ can be calculated as follows:

$$\nu_P(x) = \mathrm{NM}(\mu_P(x)) = (1-\mu_P(x)^{\beta})^{\frac{1}{\beta}},\ \beta > 0 \tag{9}$$

The intuitionistic fuzzy distance between ith element $\mathbf{x}_j^{\mathrm{IFS}} = \langle\mu(x_j), \nu(x_j), \pi(x_j)\rangle$ of an IFS X^{IFS} and jth element $\mathbf{x}_j^{\mathrm{IFS}} = \langle\mu(x_j), \nu(x_j), \pi(x_j)\rangle$ of an IFS X^{IFS} can be calculated as (Szmidt and Kacprzyk 2000):

$$\|\mathbf{x}_i^{\mathrm{IFS}} - \mathbf{x}_j^{\mathrm{IFS}}\| = \sqrt{\frac{1}{2}((\mu(x_i)-\mu(x_j))^2+(\nu(x_i)-\nu(x_j))^2+(\pi(x_i)-\pi(x_j))^2)} \tag{10}$$

2.2 Intuitionistic fuzzy representation of data

Intuitionistic fuzzy generator defined above is used to construct the intuitionistic fuzzy data for clustering (Vlachos and Sergiadis 2007). Let X be the set of data points having N elements and x_j represents the value of jth data point in set X where $j \in 1, 2, \ldots N$. These data values are normalized to [0 1] for creating the IFSs corresponding to a given set of data points. Let X^{IFS} be the IFS defined on set X with jth element $\mathbf{x}_j^{\mathrm{IFS}}$ as IFS element $\langle\mu_X(x_j), \nu_X(x_j), \pi_X(x_j)\rangle$ corresponding to jth data point with value x_j, where $\mu_X(x_j)$ is the normalized value and $\nu_X(x_j)$ is calculated using one of the negation functions described above. $\mathbf{U}$ represents the fuzzy partition matrix defined as $\mathbf{U} = \{\mu_i(x_j) : \mu_i(x_j) \in [0,1]; \sum_{i=1}^{c}\mu_i(x_j) = 1;\ \forall j \in \{1, 2, \ldots N\}$ and $i \in \{1, 2, \ldots c\}\}$ which divides the set of data points into c equivalence classes with each class i representing the membership value $\mu_i(x_j)$ of jth data point $\forall j \in 1, 2, \ldots N$. $\nu_i(x_j)$ and $\pi_i(x_j)$ represent the non-membership values and hesitation degree for jth data point in ith class or cluster, respectively. For simplicity, in rest of the paper, $\mu_i(x_j)$, $\nu_i(x_j)$ and $\pi_i(x_j)$ are denoted as μ_{ij}, ν_{ij} and π_{ij}, respectively.

2.3 Kernel methods

In recent past, many kernel-based fuzzy clustering methods have been proposed in the literature (Graves and Pedrycz 2010). The key feature of kernel-based methods is to perform a nonlinear mapping ϕ from low-dimensional feature space to higher-dimensional feature space (also known as kernel space) (Muller et al. 2001). According to Cover's theorem (Cover 1965), this nonlinear mapping ϕ makes the data points linearly separable in kernel space. The advantage of using kernel method is the fact that the dot product in the kernel space can be denoted with the help of Mercer's kernel (Mercer 1909) as:

$$K(x_i, x_j) = \phi(x_i)^{\mathrm{T}}\phi(x_j) \tag{11}$$

where x_i represent data in original space, $\phi(x_i)$ denotes the data in higher-dimensional space. Thus, it can be observed from (11) that the distance in the higher-dimensional space (kernel space) need not be explicitly computed as it can be replaced by a Mercer kernel function (typically referred to as a kernel trick) which reduces the computational complexity and save the computation time. The linear kernel function, polynomial kernel function of degree p, sigmoid kernel function and the Gaussian radial basis function (RBF) kernel function are few important kernel functions, which are given, respectively, as:

$$K(x_i, x_j) = x_j^T x_i$$
$$K(x_i, x_j) = (1 + x_j^T x_i)^p$$
$$K(x_i, x_j) = \tanh(\alpha(x_j^T x_i) + \beta)$$
$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{\sigma^2}\right)$$

Let elements $\mathbf{x}_i^{\mathrm{IFS}}$ and $\mathbf{x}_j^{\mathrm{IFS}}$ mapped to higher-dimensional space through a nonlinear transform ϕ then Euclidean distance in kernel space can be derived as follows:

$$\|\phi(\mathbf{x}_i^{\mathrm{IFS}}) - \phi(\mathbf{x}_j^{\mathrm{IFS}})\|^2 = \|\phi(\mathbf{x}_i^{\mathrm{IFS}}) - \phi(\mathbf{x}_j^{\mathrm{IFS}})\|^{\mathrm{T}}\|\phi(\mathbf{x}_i^{\mathrm{IFS}}) - \phi(\mathbf{x}_j^{\mathrm{IFS}})\|$$
$$\|\phi(\mathbf{x}_i^{\mathrm{IFS}}) - \phi(\mathbf{x}_j^{\mathrm{IFS}})\|^2 = \phi(\mathbf{x}_i^{\mathrm{IFS}})^T\phi(\mathbf{x}_i^{\mathrm{IFS}}) + \phi(\mathbf{x}_j^{\mathrm{IFS}})^{\mathrm{T}}\phi(\mathbf{x}_j^{\mathrm{IFS}}) - 2\phi(\mathbf{x}_i^{\mathrm{IFS}})\phi(\mathbf{x}_j^{\mathrm{IFS}})$$

and according to (11) the above equation can be written as:

$$\|\phi(\mathbf{x}_i^{\mathrm{IFS}}) - \phi(\mathbf{x}_j^{\mathrm{IFS}})\|^2 = K(\mathbf{x}_i^{\mathrm{IFS}}, \mathbf{x}_i^{\mathrm{IFS}}) + K(\mathbf{x}_j^{\mathrm{IFS}}, \mathbf{x}_j^{\mathrm{IFS}}) - 2K(\mathbf{x}_i^{\mathrm{IFS}}, \mathbf{x}_j^{\mathrm{IFS}}) \tag{12}$$

In case of Gaussian RBF kernel, this can be rewritten as:

$$\|\phi(\mathbf{x}_i^{\mathrm{IFS}}) - \phi(\mathbf{x}_j^{\mathrm{IFS}})\|^2 = 2(1 - K(\mathbf{x}_i^{\mathrm{IFS}}, \mathbf{x}_j^{\mathrm{IFS}})) \tag{13}$$

Similarly the Euclidean distance for other kernel space can be calculated after replacing the corresponding kernel function in (12).

2.4 Intuitionistic fuzzy c-means (IFCM) algorithm

Intuitionistic fuzzy c-means (IFCM) algorithm (Xu and Wu 2010) is a variant of fuzzy c-means algorithm use to cluster data represented in form of IFSs. It incorporates the advantage of intuitionistic fuzzy sets for handling uncertainty which arises due to imprecise information. The optimization problem of IFCM is defined as:

$$\min J_m(\mathbf{U}, \mathbf{V}^{\mathrm{IFS}} : \mathbf{X}^{\mathrm{IFS}}) = \sum_{i=1}^{c}\sum_{j=1}^{N} \mu_{ij}^m \|\mathbf{x}_j^{\mathrm{IFS}} - \mathbf{v}_i^{\mathrm{IFS}}\|^2$$
$$\text{subject to} \sum_{i=1}^{c} \mu_{ij} = 1,\ 1 \le j \le N \tag{14}$$

where m represents fuzzifier factor. The solution of the above optimization problem (14) can be obtained using Lagrange method of undetermined multiplier which is given as:

$$\mu_{ij} = \left\{ \sum_{k=1}^{c} \left(\frac{\|\mathbf{x}_j^{\mathrm{IFS}} - \mathbf{v}_i^{\mathrm{IFS}}\|^2}{\|\mathbf{x}_j^{\mathrm{IFS}} - \mathbf{v}_k^{\mathrm{IFS}}\|^2} \right)^{\frac{1}{m-1}} \right\}^{-1} \tag{15}$$

$$\mu_V(v_i) = \frac{\sum_{j=1}^{N} \mu_{ij}^m \mu_X(x_j)}{\sum_{j=1}^{N} \mu_{ij}^m},\ 1 \le i \le c \tag{16}$$

$$v_V(v_i) = \frac{\sum_{j=1}^{N} \mu_{ij}^m v_X(x_j)}{\sum_{j=1}^{N} \mu_{ij}^m},\ 1 \le i \le c \tag{17}$$

$$\pi_V(v_i) = \frac{\sum_{j=1}^{N} \mu_{ij}^m \pi_X(x_j)}{\sum_{j=1}^{N} \mu_{ij}^m},\ 1 \le i \le c \tag{18}$$

The outline of IFCM algorithm is depicted in Algorithm 1 (Xu and Wu 2010):

Algorithm 1 IFCM algorithm

Require: Set values for the number of clusters c, the fuzzifier factor $m > 1$ and the stopping criterion value ε.
1: Initialize randomly the fuzzy partition matrix $U^{(1)} = \{\mu_{ij}^1\}_{c\times N}$
2: $k \leftarrow 1$
3: **repeat**
4: Update cluster centers $\mathbf{v}_i^{IFS^{(k)}}$ using (16), (17) and (18)
5: Update partition matrix $U^{(k+1)} = \{\mu_{ij}^{k+1}\}_{c\times N}$ using (15)
6: $k \leftarrow k + 1$
7: **until** $||U^{(k+1)} - U^{(k)}|| < \varepsilon$
8: **return** $\mathbf{v}_i$ the centers of clusters and the membership degrees $\mathbf{U} = \{\mu_{ij}\}_{c\times N}$.

2.5 Fuzzy entropy clustering (FEC) algorithm

Fuzzy entropy clustering (FEC) (Tran and Wagner 2000) is an alternative hard c-means algorithm, which utilizes the advantage of entropy. The optimization problem of FEC is defined as difference of the sum of square error and the fuzzy entropy multiplied by a weight factor which denotes the degree of fuzzy entropy. The optimization problem of FEC algorithm is described as (Tran and Wagner 2000):

$$\min J_\lambda(\mathbf{U}, \mathbf{V} : \mathbf{X}) = \sum_{i=1}^{c}\sum_{j=1}^{N} \mu_{ij} \|\mathbf{x}_j - \mathbf{v}_i\|^2 + \lambda \sum_{i=1}^{c}\sum_{j=1}^{N} \mu_{ij} \log \mu_{ij}$$
$$\text{subject to} \sum_{i=1}^{c} \mu_{ij} = 1,\ 1 \le j \le N \tag{19}$$

where $\mathbf{x}_j$, $j = 1, 2, \ldots, N$ are feature vectors of data points, c is the number of clusters, λ is the degree of fuzzy entropy, μ_{ij} is the membership degree of the jth data points to the ith cluster, U is fuzzy partition matrix, $V = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_c\}$ is the set of centroids of clusters, and $\|\mathbf{x}_j - \mathbf{v}_i\|$ is Euclidean distance between the jth data point to the ith cluster centroid.

The second term in the optimization problem (19) is known as fuzzy entropy function, which measures the average degree of non-membership of fuzzy sets. Using Lagrange method of undetermined multiplier, the optimization problem (19) can be solved. The membership and centroid of FEC method are, respectively, given as:

$$\mu_{ij} = \left\{ \sum_{k=1}^{c} \left(\frac{\exp(\|\mathbf{x}_j - \mathbf{v}_i\|^2)}{\exp(\|\mathbf{x}_j - \mathbf{v}_k\|^2)} \right)^{\frac{1}{\lambda}} \right\}^{-1} \tag{20}$$

$$v_i = \frac{\sum_{j=1}^{N} \mu_{ij} x_j}{\sum_{j=1}^{N} \mu_{ij}} \tag{21}$$

The outline of FEC algorithm is depicted in Algorithm 2 (Tran and Wagner 2000):

Algorithm 2 FEC algorithm

Require: Set values for the number of clusters c, the fuzzy entropy factor $\lambda > 0$ and the stopping criterion value ε.

1: Initialize randomly the fuzzy partition matrix $U^{(1)} = \{\mu_{ij}^1\}_{c\times N}$
2: $k \leftarrow 1$
3: **repeat**
4: Update cluster centers $v_i^{(k)}$ using (21)
5: Update partition matrix $U^{(k+1)} = \{\mu_{ij}^{k+1}\}_{c\times N}$ using (20)
6: $k \leftarrow k + 1$
7: **until** $||U^{(k+1)} - U^{(k)}|| < \varepsilon$
8: **return** $\mathbf{v}_i$ the centers of clusters and the membership degrees $\mathbf{U} = \{\mu_{ij}\}_{c\times N}$.

3 Proposed kernel intuitionistic fuzzy entropy c-means (KIFECM) method

In this work, a new kernel intuitionistic fuzzy entropy c-means (KIFECM) algorithm is proposed for handling the drawbacks of related works (Hai-Jun et al. 2011; Tran and Wagner 2000) in the literature. The objective function of the proposed approach consists of three terms. The first term is the total within clustering error after the transformation of data points in higher-dimensional space to increase the representation capability with the help of nonlinear transformation. The second term $\sum_{i=1}^{c}\sum_{j=1}^{N}(\mu_{ij}\log\mu_{ij} - \mu_{ij})$, a monotone decreasing function in μ_{ij} forces μ_{ij} to be maximum to increase the chances of getting non-trivial solution. The third term is the negative of $-\sum_{i=1}^{c}\sum_{j=1}^{N}\mu_{ij}\log\mu_{ij}$ with a fuzzy entropy factor λ to obtain a good partition of IFSs which will increase the confidence over obtained cluster prototype. The optimization problem of the proposed method will reveal nonlinear structures present in a robust way as the distance is measured in higher-dimensional space using kernel trick without increasing the computational complexity. The optimization problem of the proposed KIFECM method can be given as:

$$\begin{aligned}\min J_\lambda^\phi(\mathbf{U}, \mathbf{V}^{\text{IFS}} : \mathbf{X}^{\text{IFS}}) &= \sum_{i=1}^{c}\sum_{j=1}^{N}\mu_{ij}\|\phi(\mathbf{x}_j^{\text{IFS}}) \\ &\quad - \phi(\mathbf{v}_i^{\text{IFS}})\|^2 \\ &\quad + \frac{\sigma^2}{c}\sum_{i=1}^{c}\sum_{j=1}^{N}(\mu_{ij}\log\mu_{ij} - \mu_{ij}) + \lambda\sum_{i=1}^{c}\sum_{j=1}^{N}\mu_{ij}\log\mu_{ij} \\ \text{subject to}\sum_{i=1}^{c}\mu_{ij} &= 1,\ 1 \le j \le N \end{aligned} \tag{22}$$

where ϕ is mapping used to transform samples from low-dimensional to higher-dimensional space and σ^2 is the measure of degree of separation of the dataset which can be defined as:

$$\sigma^2 = \frac{1}{N}\sum_{j=1}^{N}\|\mathbf{x}_j^{\text{IFS}} - \bar{\mathbf{x}}^{\text{IFS}}\|^2 \text{ with } \bar{\mathbf{x}}^{\text{IFS}} = \frac{1}{N}\sum_{j=1}^{N}\mathbf{x}_j^{\text{IFS}}.$$

For Gaussian radial basis kernel function, (22) using (13) can be rewritten as:

$$\begin{aligned}\min J_\lambda^\phi(\mathbf{U}, \mathbf{V}^{\text{IFS}} : \mathbf{X}^{\text{IFS}}) &= \sum_{i=1}^{c}\sum_{j=1}^{N}\mu_{ij}(1 - K(\mathbf{x}_j^{\text{IFS}}, \mathbf{v}_i^{\text{IFS}})) \\ &\quad + \frac{\sigma^2}{c}\sum_{i=1}^{c}\sum_{j=1}^{N}(\mu_{ij}\log\mu_{ij} - \mu_{ij}) + \lambda\sum_{i=1}^{c}\sum_{j=1}^{N}\mu_{ij}\log\mu_{ij} \\ \text{subject to}\sum_{i=1}^{c}\mu_{ij} &= 1,\ 1 \le j \le N \end{aligned} \tag{23}$$

where $K(\mathbf{x}_j^{\text{IFS}}, \mathbf{v}_i^{\text{IFS}}) = \exp(-\frac{\|\mathbf{x}_j^{\text{IFS}} - \mathbf{v}_i^{\text{IFS}}\|^2}{\sigma^2})$, cluster centers $\mathbf{v}_i^{\text{IFS}} = \langle\mu_V(v_i), \nu_V(v_i), \pi_V(v_i)\rangle \quad \forall i = 1, 2, \ldots, c$ and $\mathbf{x}_j^{\text{IFS}} = \langle\mu_X(x_j), \nu_X(x_j), \pi_X(x_j)\rangle\ \forall j = 1, 2, \ldots, N$.

The solution of the above optimization problem, using Lagrange method of undetermined multiplier, is included in "Appendix A". The membership value μ_{ij} and cluster centroid $\mathbf{v}_i^{\text{IFS}} = \langle\mu_V(v_i), \nu_V(v_i), \pi_V(v_i)\rangle$ can, respectively, be given as:

$$\mu_{ij} = \frac{\exp\left(-\frac{c[(1-K(\mathbf{x}_j^{\text{IFS}}, \mathbf{v}_i^{\text{IFS}}))+\lambda]}{(\sigma^2+c\lambda)}\right)}{\sum_{k=1}^{c}\exp\left(-\frac{c[(1-K(\mathbf{x}_j^{\text{IFS}}, \mathbf{v}_k^{\text{IFS}}))+\lambda]}{(\sigma^2+c\lambda)}\right)} \tag{24}$$

above formulation for μ_{ij} can be rewritten as:

$$\mu_{ij} = \left\{\sum_{k=1}^{c}\left(\frac{\exp(1 - K(\mathbf{x}_j^{\text{IFS}}, \mathbf{v}_i^{\text{IFS}}))}{\exp(1 - K(\mathbf{x}_j^{\text{IFS}}, \mathbf{v}_k^{\text{IFS}}))}\right)^{\frac{1}{\frac{\sigma^2}{c}+\lambda}}\right\}^{-1} \tag{25}$$

$$\mu_V(v_i) = \frac{\sum_{j=1}^{N}\mu_{ij}(K(\mathbf{x}_j^{\text{IFS}}, \mathbf{v}_i^{\text{IFS}})\mu_X(x_j))}{\sum_{j=1}^{N}\mu_{ij}(K(\mathbf{x}_j^{\text{IFS}}, \mathbf{v}_i^{\text{IFS}}))} \tag{26}$$

$$\nu_V(v_i) = \frac{\sum_{j=1}^{N}\mu_{ij}(K(\mathbf{x}_j^{\text{IFS}}, \mathbf{v}_i^{\text{IFS}})\nu_X(x_j))}{\sum_{j=1}^{N}\mu_{ij}(K(\mathbf{x}_j^{\text{IFS}}, \mathbf{v}_i^{\text{IFS}}))} \tag{27}$$

$$\pi_V(v_i) = \frac{\sum_{j=1}^{N}\mu_{ij}(K(\mathbf{x}_j^{\text{IFS}}, \mathbf{v}_i^{\text{IFS}})\pi_X(x_j))}{\sum_{j=1}^{N}\mu_{ij}(K(\mathbf{x}_j^{\text{IFS}}, \mathbf{v}_i^{\text{IFS}}))} \tag{28}$$

Equation (25) asserts that μ_{ij}, the membership value of jth data point to the ith cluster, is a function of distance in kernel space with associated parameter values. These parameter values include fuzzy negation parameter and fuzzy entropy factor which requires tuning as is discussed in Sect. 4.2. It is easy to follow that the membership value is analogous to (15) and (20) and it lies in [0, 1] where λ works as a fuzzy entropy factor.

Cluster centroids, $\mathbf{v}_i^{\mathrm{IFS}} = \langle \mu_V(v_i), \nu_V(v_i), \pi_V(v_i) \rangle$, given by (26), (27) and (28) are endowed with additional weight that reads $K(\mathbf{x}_j^{\mathrm{IFS}}, \mathbf{v}_i^{\mathrm{IFS}})$. It measures the similarity between x_j^{IFS} and v_i^{IFS}. Moreover, for noisy and outlier data points the value of $K(\mathbf{x}_j^{\mathrm{IFS}}, \mathbf{v}_i^{\mathrm{IFS}})$ becomes very small and the weighted sum get suppressed which in turn ensures the robustness of the method.

Algorithm 3 shows the steps of the iterative procedure for obtaining the partition matrix and cluster centroid using the proposed method.

Algorithm 3 Proposed KIFECM algorithm.

Require: Set values for the number of clusters c, the degree of fuzzy entropy λ, fuzzy negation parameter β, σ, and the stopping criterion value ε.

1: Calculate $\mathbf{x}_j^{\mathrm{IFS}} = \langle \mu_X(x_j), \nu_X(x_j), \pi_X(x_j) \rangle \; \forall j \in \{1, 2, \ldots, N\}$ as defined in Sec. 2.2
2: Initialize the cluster centers $\mathbf{v}_i^{\mathrm{IFS}(1)} = \langle \mu_V(x_j)^{(1)}, \nu_V(x_j)^{(1)}, \pi_V(x_j)^{(1)} \rangle$ randomly
3: $k \leftarrow 1$
4: **repeat**
5: Calculate the updated fuzzy partition matrix $U^{(k)} = \{\mu_{ij}^{(k)}\}_{c \times N}$ using (24)
6: Update the cluster centers $\mathbf{v}_i^{\mathrm{IFS}(k+1)} = \langle \mu_V(x_j)^{(k+1)}, \nu_V(x_j)^{(k+1)}, \pi_V(x_j)^{(k+1)} \rangle$ using (26), (27) and (28)
7: $k \leftarrow k + 1$
8: **until** $\| \mathbf{v}_i^{\mathrm{IFS}(k+1)} - \mathbf{v}_i^{\mathrm{IFS}(k)} \| < \varepsilon$
9: **return** $\mathbf{v}_i^{\mathrm{IFS}}$ the centers of clusters and the membership degrees $\mathbf{U} = \{\mu_{ij}\}_{c \times N}$.

The key features of the proposed KIFECM method reads:

- It reveals the inherent non-Euclidean structures by incorporating the kernel space transformation and that too without increasing the computational complexity.
- It force the membership value to be maximum so as to increase the chances of getting a non-trivial solution.
- It helps obtaining a good partition of IFSs by incorporation of fuzzy entropy function.
- It produces clusters which are robust to noise and outlier.

4 Experimental setup and results

To check the effectiveness of the proposed KIFECM method over related works FEC, FCM, IFCM, UPCA, PIFECM and KFEC (Pham et al. 2018), experiments are performed on two synthetic datasets and three publicly available datasets. The performance measures used for comparison are given in Table 1 (Anbeek et al. 2004; Bezdek 1981; Chuang et al. 2006; Qiu et al. 2013; Vovk et al. 2007). The descriptions of the variables in Table 1 are as follows: c, N, μ_{ij} and m represent the number of cluster, the total number of data points, the membership of jth data point in ith cluster and the fuzzifier factor, respectively. X_i and Y_i denote the pixels belonging to the manual segmented image (ground truth) and experimental segmented image corresponding to the ith region, respectively. $|X_i|$ denotes the cardinality of X_i.

Table 1 List of the performance measures

Name of performance measures	Formula
Average segmentation accuracy	$\mathrm{ASA} = \sum_{i=1}^{c} \frac{\lvert X_i \cap Y_i \rvert}{\sum_{j=1}^{c} \lvert X_j \rvert}$
Dice Score	$\mathrm{DS} = \frac{2\lvert X_i \cap Y_i \rvert}{\lvert X_i \rvert + \lvert Y_i \rvert}$
Jaccard Score	$\mathrm{JS} = \frac{\lvert X_i \cap Y_i \rvert}{\lvert X_i \cup Y_i \rvert}$
False positive ratio relative to ground truth	$\mathrm{FPR} = \frac{\lvert Y_i \rvert - \lvert X_i \cap Y_i \rvert}{\lvert X_i \rvert}$
False negative ratio relative to ground truth	$\mathrm{FNR} = \frac{\lvert X_i \rvert - \lvert X_i \cap Y_i \rvert}{\lvert X_i \rvert}$
Partition coefficient	$\mathrm{PC} = \frac{1}{N} \sum_{i=1}^{c} \sum_{j=1}^{N} \mu_{ij}^m$
Partition entropy	$\mathrm{PE} = -\frac{1}{N} \sum_{i=1}^{c} \sum_{j=1}^{N} \mu_{ij} \log \mu_{ij}$

4.1 Datasets

Two new synthetic datasets for experimentation have been created. The first one is 2-D data points which is generated around vertices of an equilateral triangle, whose each side is of 6 length unit, to be the actual cluster centers. Around this assumed cluster centers, we have generated the 2-D data points inside a circular region with radius 3. We have generated the 300 data points around each of these vertices. We refer this dataset as triangle dataset. The other synthetic dataset is a gray image which has four regions with intensities 4, 76, 148 and 201 (See Fig. 4). Other than these synthetic datasets, we have utilized three publicly available datasets. The popular Iris dataset (Bezdek et al. 1999; Fisher 1936) is used for evaluating the performance of the proposed method which contains 150 data points of three classes (50 each class) with four features.

BrainWeb simulated MRI brain volumes, a publicly available dataset (Cocosco et al. 1997) contained simulated T1-weighted MRI images with different levels of noise (1%, 3%, 5%, 7% and 9%) and intensity inhomogeneity or intensity non-uniformity (INU) (0%, 20% and 40%) of resolution $1 \times 1 \times 1\,\mathrm{mm}^3$ with $181 \times 217 \times 181$ dimension.

Another publicly available real MRI brain images have been acquired from the Internet Brain Segmentation Repository (IBSR) (available: https://www.nitrc.org/projects/ibsr) which have the ground truth data along with it. For all the MRI images, the brain extraction tool (available: http://www.fmrib.ox.ac.uk/fsl/.) is utilized for scull striping.

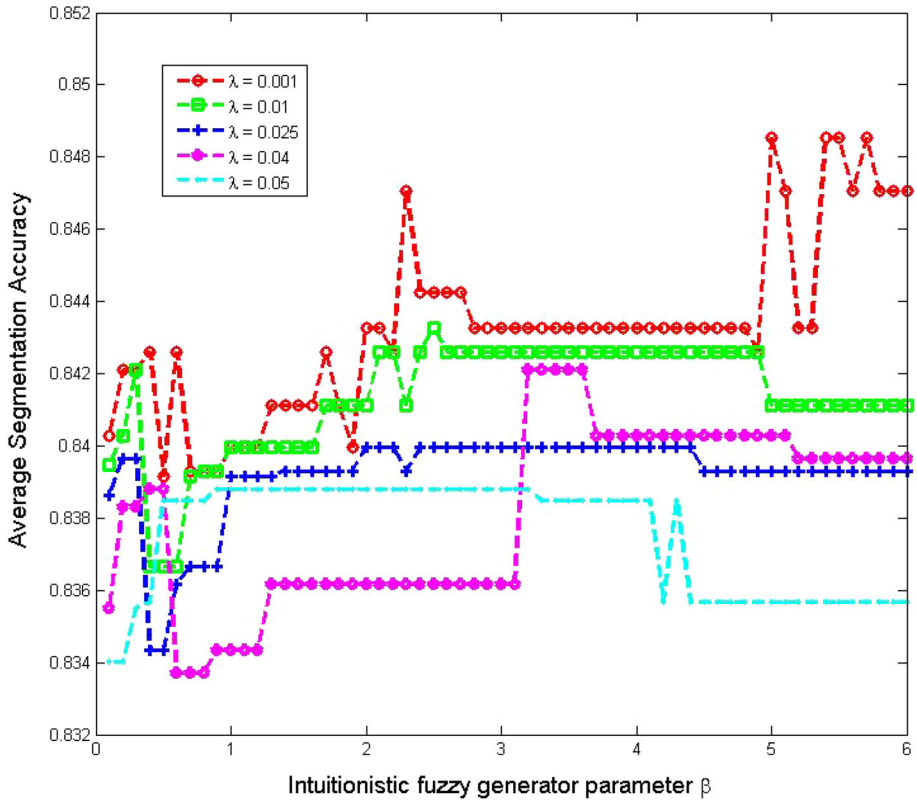


Fig. 1 Performance of the proposed KIFECM method with SNF in terms of Average Segmentation Accuracy by fixing fuzzy entropy factor λ and varying intuitionistic fuzzy generator parameter β

4.2 Parameter selection

The proposed KIFECM method involves two parameters λ and intuitionistic fuzzy negation parameter β, which have significant influence on the solution of its optimization problem, i.e., partition matrix and the cluster prototypes, according to (24), (26), (27) and (28), thereby affecting the clustering accuracy. To find the optimal parameters corresponding to each dataset, experiments have been done in 3 steps. First, the performance of the proposed method in terms of ASA is done by fixing the fuzzy entropy parameter λ and varying the intuitionistic fuzzy negation parameter β. In second step, same procedure is applied by fixing the intuitionistic fuzzy negation parameter β and varying the fuzzy entropy parameter λ. Finally, a set of optimal values corresponding to a selected range of these parameters is obtained based on the performance of the method in terms of ASA. Similar procedure is followed for obtaining the optimal values of the parameters associated with other related works for comparing the proposed method. In this work, the range of values for parameters λ and β are set to [0 0.1] and [0 10], respectively.

Figure 1 shows the performance of the proposed KIFECM method with SNF in terms of ASA by fixing fuzzy entropy factor λ and varying Sugeno's Intuitionistic fuzzy generator parameter on IBSR case 13_3 MRI data on 134th slice. It can be noted that the ASA achieved higher value when λ and β are in the range of [0.001 0.01] and [5 6], respectively, corresponding to the proposed KIFECM method with SNF. Table 2 shows the optimal value of the parameter associated with different related works and the proposed method for different datasets. The stopping criterion value ε is set to 0.001 for all the methods.

4.3 Experimental results and discussion

In this subsection, a detailed discussion and comparison of the performance of the proposed KIFECM method with SNF and YNF and existing methods such as FEC, FCM, IFCM, UPCA, PTFECM and KFEC (Pham et al. 2018) have been given in terms of aforementioned performance measures on different datasets separately. The result in boldface shows the best performing method.

Results on synthetic triangle dataset

The proposed KIFECM method is compared for SNF and YNF with FEC, FCM, IFCM, UPCA, PTFECM and KFEC (Pham et al. 2018) on a synthetic triangle dataset. Figure 2 reveals the original data point of the dataset with given ground truth given in different colors. Moreover, Fig. 3a–h shows, respectively, the qualitative performance of FEC, FCM, IFCM, UPCA, PTFECM, KFEC (Pham et al. 2018) and the proposed KIFECM method with SNF and YNF. It

Table 2 Value of parameter used for experiment for different datasets

Method\dataset	Triangle dataset	Iris dataset	Synthetic gray image data	BrainWeb simulated dataset	IBSR dataset
FEC	c=3; λ=5.1	c=3; λ=1.44	c=4; λ=450	c=4; λ=550	c=4; λ=450
FCM	m=2; c=3	m=2; c=3	m=2; c=4	m=2; c=4	m=2; c=4
IFCM	m=2; c=3; β=0.09	m=2; c=3; β=0.93	m=2; c=4; β=0.93	m=2; c=4; β=1.8	m=2; c=4; β=0.8
UPCA	m=2; c=3	m=2; c=3	m=2; c=4	m=3; c=4	m=10; c=4
PTFECM	m=2; c=3; λ=2.4	m=2; c=3; λ=0.01	m=2; c=4; λ=30	m=6; c=4; λ=0.8	m=9; c=4; λ=15
KFEC (Pham et al. 2018)	c=3; λ=0.3	c=3; λ=0.3	c=4; λ=0.01	c=4; λ=0.03	c=4; λ=0.03
KIFECM					
SNF	c=3; β=6.5; λ=0.1	c=3; β=0.9; λ=0.13	c=4; β=0.6; λ=0.02	c=4; β=2.5; λ=0.005	c=4; β=5.5; λ=0.001
YNF	c=3; β=0.5; λ=0.04	c=3; β=0.63; λ=0.23	c=4; β=0.63; λ=0.02	c=4; β=0.63; λ=0.005	c=4; β=2.5; λ=0.01

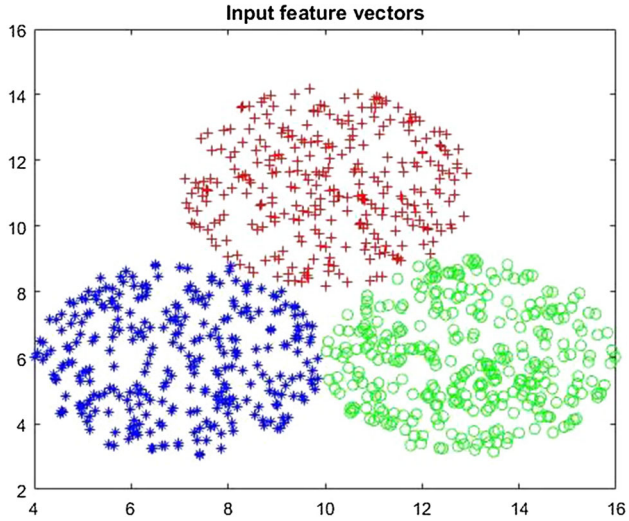


Fig. 2 Triangle dataset

is easy to follow that the proposed method outperforms the other related works. It can also be observed from Fig. 3 that all the methods are performing well in assigning the inner data points to the correct cluster. But for boundary data points, the others methods are unable to assigning data points to correct cluster except the proposed KIFECM method. The proposed KIFECM method exploits the advantage of the kernel function which enables us to find nonlinear structures. Further, the quantitative evaluation is also performed on this dataset in terms of PC and PE and class-wise clustering accuracy. Tables 3 and 4 contain the quantitative values corresponding to PC and PE and class-wise clustering accuracy, respectively, for FEC, FCM, IFCM, UPCA, PTFECM and the proposed KIFECM method for SNF and YNF. From Table 3, it can be seen that the proposed KIFECM method for both SNF and YNF achieved the highest PC value and lowest clustering entropy PE in comparison with other methods, which indicates the superior clustering capability of the proposed method in comparison with other methods. The same is true for the class-wise clustering accuracy as shown in Table 4.

Results on synthetic gray image dataset

As discussed in the dataset section, we have synthesize gray image shown in Fig. 4 which contains four different intensity regions. The performance measures in terms of PC, PE and ASA corresponding to the FEC, FCM, IFCM, UPCA, PTFECM, KFEC (Pham et al. 2018) and the proposed KIFECM method for SNF and YNF, for the gray images corrupted with Gaussian white noise with mean and variance $\mu = 0, \sigma^2 = 0.001$, $\mu = 0, \sigma^2 = 0.002$ and $\mu = 0, \sigma^2 = 0.003$ are shown in Table 5, 6 and 7, respectively. Tables 5, 6 and 7 suggest that increase in the level of noise deteriorates the performance of other methods to a greater extent in terms of PC, PE and ASA when compared to IFCM, while the proposed KIFECM method is robustness toward the noise as desired. Also, it can be noted that the proposed KIFECM method for both SNF and YNF outperforms the other methods on each noise corrupted gray images separately.

Results on Iris dataset

Iris dataset is one of the popular real datasets which is widely utilized for validating the performance of several supervised and unsupervised learning technique. So, this dataset is used to obtain the performance of the proposed method and other related works. Tables 8 and 9 show the quantitative values of PC and PE and class-wise accuracy obtained from FEC, FCM, IFCM, UPCA, PTFECM, KFEC (Pham et al. 2018) and the proposed KIFECM method for SNF and YNF. It can be observed from Table 9 that the performance of the proposed method is well enough in terms of class-wise accuracy due to the exploitation of kernel. However, at the same time, it can be noted from Table 8 that the value of partition coeffi-

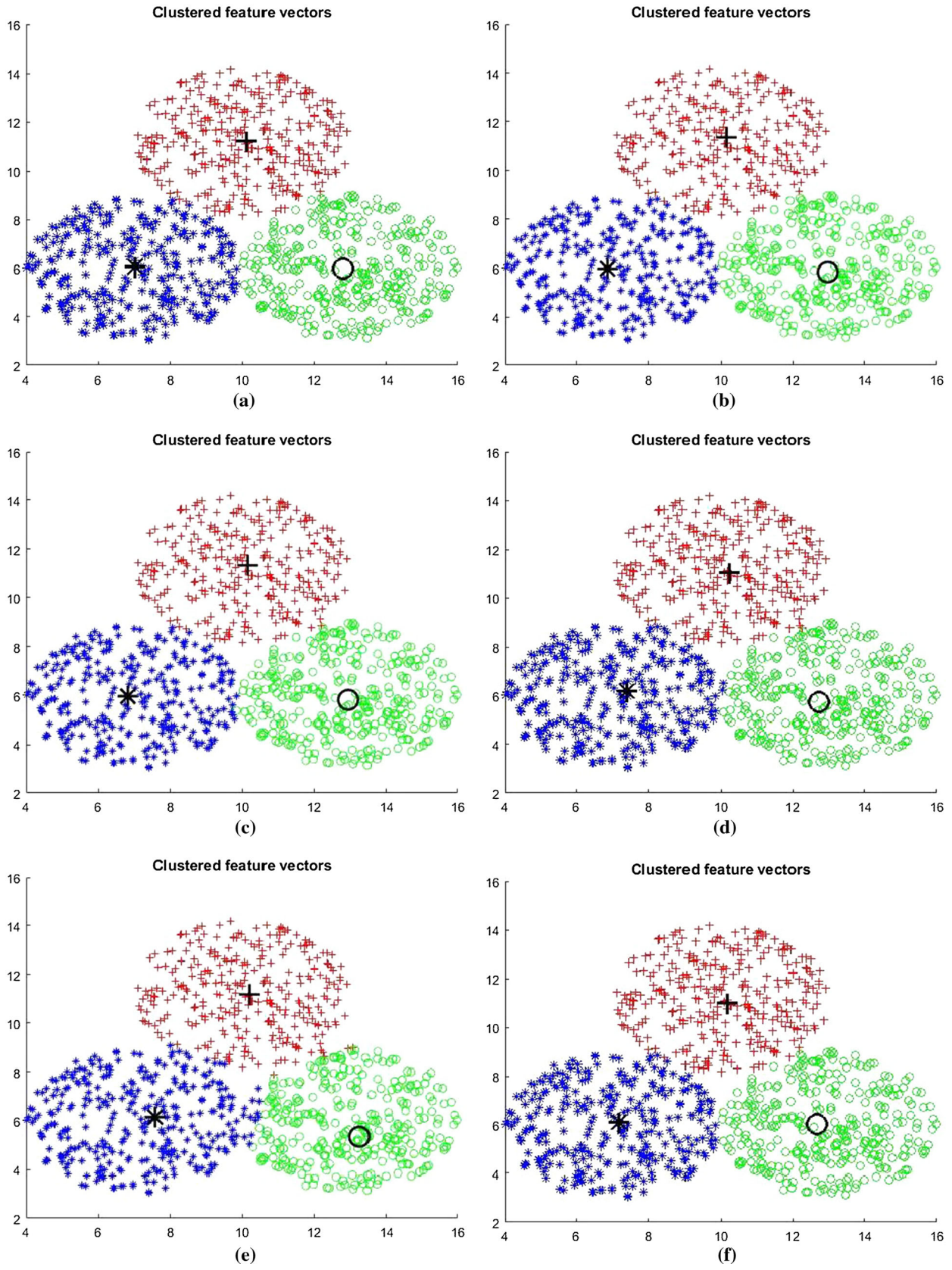


Fig. 3 Comparison of clustering methods on triangle dataset **a** FEC, **b** FCM, **c** IFCM **d** PTFECM, **e** UPCA **f** KFEC (Pham et al. 2018),**g** KIFECM with SNF, **h** KIFECM with YNF

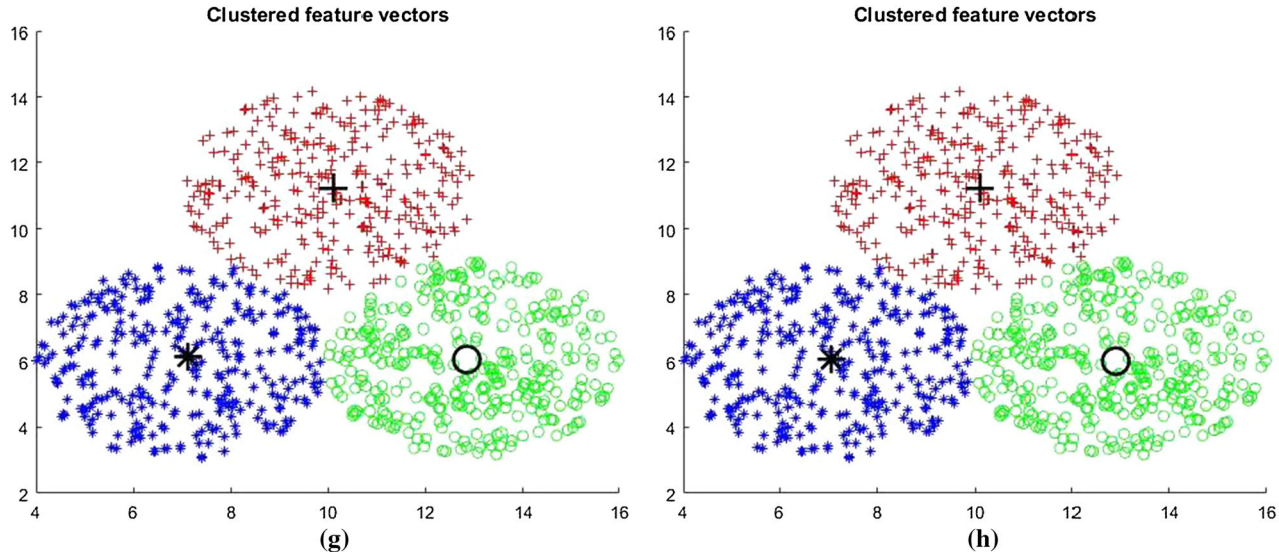


Fig. 3 continued

Table 3 Cluster validity values for synthetic triangle dataset

	FEC	FCM	IFCM	UPCA	PTFECM	KFEC (Pham et al. 2018)	KIFECM	
							SNF	YNF
PC	0.9028	0.6887	0.6878	0.2648	0.9276	0.6382	0.9339	**0.9757**
PE	0.1728	0.5729	0.5742	0.3477	0.1270	0.6598	0.1142	**0.0420**

Table 4 Class-wise accuracy for synthetic triangle dataset

	FEC	FCM	IFCM	UPCA	PTFECM	KFEC (Pham et al. 2018)	KIFECM	
							SNF	YNF
Class1	0.9989	0.9989	**1.0000**	0.9778	0.9944	0.9978	**1.0000**	**1.0000**
Class2	0.9978	0.9978	0.9989	0.9778	0.9944	0.9967	**1.0000**	**1.0000**
Class3	0.9989	0.9989	0.9989	**1.0000**	**1.0000**	0.9989	**1.0000**	**1.0000**

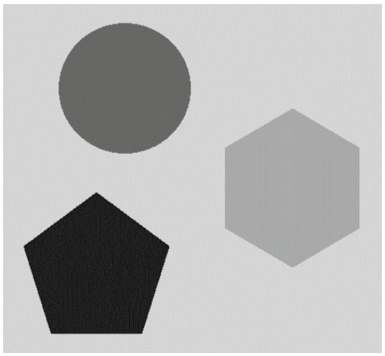


Fig. 4 Synthetic gray image dataset

cient PC is maximum and partition entropy PC is minimum for the proposed method compared to other related methods which indicate the superior clustering capability of the proposed KIFECM method in comparison with other methods.

Results on simulated brain images

The experiments have been done for comparing the performance of the proposed KIFECM method with SNF and YNF on Simulated Brain MRI dataset. Tables 10, 11, 12 and 13 show the comparison of other existing methods with FEC, FCM, IFCM, UPCA, PTFECM and KFEC (Pham et al. 2018) on Simulated Brain MRI dataset corresponding to MR images with INU 0%, 20% and 40% with noise level 7% and 9% on slice 90 in terms of performance measures DS, JS, FPR and FNR, respectively, for WM, GM and CSF. It is evident from Tables 10 and 11 that the performance of the proposed KIFECM method with SNF and YNF in terms of DS and JS for WM achieves high score compared to other related methods, whereas these scores for GM and CSF in most of the cases favor the proposed method in the presence of different levels of noise and INU. From Tables 12 and 13, it can be observed that the false positive ratio and false neg-

Table 5 Result on synthetic image with Gaussian white noise ($\mu = 0, \sigma^2 = 0.001$)

	FEC	FCM	IFCM	UPCA	PTFECM	KFEC (Pham et al. 2018)	KIFECM	
							SNF	YNF
PC	0.9789	0.9458	0.9457	0.5909	0.9880	0.9981	**0.9987**	0.9980
PE	0.0671	0.1833	0.1829	0.8471	0.0344	0.0046	**0.0036**	0.0057
ASA	0.9992	**0.9996**	**0.9996**	0.6551	0.9974	0.9994	**0.9996**	0.9995

Table 6 Result on synthetic image with Gaussian white noise ($\mu = 0, \sigma^2 = 0.002$)

	FEC	FCM	IFCM	UPCA	PTFECM	KFEC (Pham et al. 2018)	KIFECM	
							SNF	YNF
PC	0.8712	0.8980	0.9000	0.5900	0.9754	0.9722	**0.9885**	0.9818
PE	0.3074	0.3068	0.3007	0.8487	0.0644	0.0493	**0.0284**	0.0447
ASA	0.9401	0.9879	**0.9897**	0.6487	0.9879	0.9870	**0.9897**	0.9860

Table 7 Result on synthetic image with Gaussian white noise ($\mu = 0, \sigma^2 = 0.003$)

	FEC	FCM	IFCM	UPCA	PTFECM	KFEC (Pham et al. 2018)	KIFECM	
							SNF	YNF
PC	0.8300	0.8483	0.8610	0.5894	0.6133	0.9610	0.9638	**0.9724**
PE	0.3898	0.4253	0.3940	0.8493	0.7799	0.0957	0.0852	**0.0651**
ASA	0.8673	0.9425	0.9600	0.6463	0.6440	0.9550	**0.9602**	0.9515

Table 8 Cluster validity values for Iris dataset

	FEC	FCM	IFCM	UPCA	PTFECM	KFEC (Pham et al. 2018)	KIFECM	
							SNF	YNF
PC	0.7880	0.7834	0.7390	0.6777	0.2605	0.5819	**0.8804**	0.8045
PE	0.3267	0.3955	0.4696	0.3776	0.3217	0.6881	**0.1820**	0.3000

Table 9 Class-wise clustering accuracy for Iris dataset

	FEC	FCM	IFCM	UPCA	PTFECM	KFEC (Pham et al. 2018)	KIFECM	
							SNF	YNF
Class1	**1.0000**	**1.0000**	**1.0000**	0.9933	**1.0000**	0.9933	**1.0000**	**1.0000**
Class2	0.9600	0.9333	0.9933	0.9467	0.9733	0.9867	**1.0000**	**1.0000**
Class3	0.9600	0.9333	0.9933	0.9533	0.9733	0.9933	**1.0000**	**1.0000**

ative ratio relative to the ground truth of different tissues are less for the proposed method compared to other related method due to the presence of INU. FPR values for WM are high in comparison with GM as the INU affects the intensity values of GM region compared to that of WM, which leads to over-segmentation for WM. Hence, the FNR values for GM are high in comparison with WM. Tables 14, 15 and 16 show the performance in terms of ASA, PC and PE for FEC, FCM, IFCM, UPCA, PTFECM, KFEC (Pham et al. 2018) and the proposed KIFECM method with SNF and YNF. Table 14 shows the comparison scores in terms of ASA which clearly indicate that the proposed method KIFECM with SNF and YNF outperforms for different levels of noise and INU. Although, the level of noise and INU increases, the performance of almost every method deteriorates, but IFCM and the proposed KIFECM have satisfactory performance in terms of DS, JS and ASA as these methods are robust to noise. At the same time, it can be observed from Tables 15 and 16 that the performance of the proposed method in terms of PC and PE values is best in comparison with other related methods which increases the confidence on the partition.

Figure 6 shows qualitative results obtained for different methods along with the proposed KIFECM with SNF and YNF row-wise corresponding to MR image (INU = 0% and noise level = 7%) on an axial slice 90 shown in Fig. 5 with ground truth corresponding to WM, GM and CSF. It can be pointed out from the red and green colored circle regions in Fig. 6 that the segmentation

Table 10 DS for BrainWeb simulated MRI

INU	Noise	FEC	FCM	IFCM	UPCA	PTFECM	KFEC (Pham et al. 2018)	KIFECM	
								SNF	YNF
WM 0	7	0.9233	0.9237	0.9264	0.9195	0.9233	0.9261	**0.9271**	**0.9271**
	9	0.8768	0.8793	0.8890	0.8612	0.8720	0.8869	**0.8909**	**0.8909**
20	7	0.9241	0.9238	0.9258	0.9179	0.9238	0.9255	**0.9266**	**0.9266**
	9	0.8790	0.8798	0.8882	0.7379	0.8877	0.8863	**0.8891**	**0.8891**
40	7	0.9057	0.9045	0.9085	0.8941	0.9069	0.9080	**0.9091**	**0.9091**
	9	0.8714	0.8702	0.8782	0.8606	0.8733	0.8770	**0.8814**	**0.8814**
GM 0	7	0.8593	0.8649	**0.8747**	0.8625	0.8718	0.8665	0.8744	0.8743
	9	0.7835	0.7994	0.8252	0.7716	0.7978	0.8091	**0.8260**	0.8257
20	7	0.8676	0.8712	**0.8782**	0.8674	0.8771	0.8724	0.8777	0.8780
	9	0.7865	0.7994	**0.8235**	0.6263	0.8059	0.8051	**0.8235**	0.8231
40	7	0.8427	0.8449	**0.8598**	0.8367	0.8541	0.8489	0.8592	0.8585
	9	0.7804	0.7909	0.8150	0.7655	0.7932	0.7990	**0.8169**	0.8165
CSF 0	7	0.7903	0.7999	0.8105	0.8040	0.8118	0.8008	0.8106	**0.8113**
	9	0.7301	0.7533	0.7950	0.7398	0.7647	0.7631	0.7972	**0.7981**
20	7	0.8172	0.8231	0.8297	0.8263	**0.8317**	0.8245	0.8292	0.8282
	9	0.7370	0.7590	0.8027	0.6512	0.7738	0.7658	0.8044	**0.8052**
40	7	0.8018	0.8050	0.8306	0.8104	0.8222	0.8092	**0.8308**	0.8291
	9	0.7368	0.7534	0.7946	0.7345	0.7637	0.7669	**0.7958**	0.7955

Table 11 JS for BrainWeb simulated MRI

INU	Noise	FEC	FCM	IFCM	UPCA	PTFECM	KFEC (Pham et al. 2018)	KIFECM	
								SNF	YNF
WM 0	7	0.8576	0.8583	0.8629	0.8510	0.8576	0.8623	**0.8641**	**0.8641**
	9	0.7807	0.7845	0.8001	0.7562	0.7731	0.7967	**0.8032**	**0.8032**
20	7	0.8590	0.8584	0.8618	0.8483	0.8584	0.8613	**0.8632**	**0.8632**
	9	0.7841	0.7854	0.7989	0.5846	0.7981	0.7958	0.8003	**0.8004**
40	7	0.8277	0.8257	0.8324	0.8085	0.8296	0.8315	**0.8333**	**0.8333**
	9	0.7721	0.7703	0.7829	0.7553	0.7751	0.7810	**0.7880**	**0.7880**
GM 0	7	0.7533	0.7620	**0.7773**	0.7583	0.7727	0.7644	0.7768	0.7766
	9	0.6441	0.6658	0.7025	0.6281	0.6636	0.6794	**0.7036**	0.7031
20	7	0.7661	0.7718	**0.7828**	0.7659	0.7811	0.7737	0.7820	0.7825
	9	0.6482	0.6658	**0.6999**	0.4559	0.6749	0.6738	**0.6999**	0.6993
40	7	0.7281	0.7314	**0.7541**	0.7192	0.7453	0.7375	0.7532	0.7521
	9	0.6398	0.6541	0.6878	0.6201	0.6573	0.6652	**0.6905**	0.6899
CSF 0	7	0.6533	0.6665	0.6814	0.6722	0.6832	0.6678	0.6815	**0.6825**
	9	0.5749	0.6042	0.6598	0.5871	0.6190	0.6169	0.6628	**0.6641**
20	7	0.6909	0.6994	**0.7090**	0.7040	0.7118	0.7014	0.7083	0.7067
	9	0.5835	0.6116	0.6705	0.4828	0.6311	0.6204	0.6728	**0.6739**
40	7	0.6691	0.6737	0.7102	0.6812	0.6981	0.6795	**0.7106**	0.7080
	9	0.5832	0.6043	0.6592	0.5804	0.6177	0.6219	**0.6608**	0.6604

results obtained using the proposed method are robust to noise to a greater extent when compared to other related works.

Table 17 shows the average computation time (50 runs on MATLAB 2016a with 8 GB RAM) taken to obtain the segmentation a brain MR image with INU 0% and noise 7% of size 181 × 217 for different methods. It can be noted that the proposed method consumes less time when compared to other methods except FEC, IFCM and KFEC (Pham et al. 2018) at the same time gives com-

Table 12 FPR for BrainWeb simulated MRI

INU	Noise	FEC	FCM	IFCM	UPCA	PTFECM	KFEC (Pham et al. 2018)	KIFECM	
								SNF	YNF
WM 0	7	0.0618	0.0620	0.0701	**0.0530**	0.0618	0.0679	0.0737	0.0737
	9	0.0684	0.0701	0.0864	**0.0517**	0.0610	0.0806	0.0926	0.0926
20	7	0.0522	0.0497	0.0581	**0.0392**	0.0497	0.0576	0.0611	0.0611
	9	0.0568	0.0578	0.0771	**0.0126**	0.0724	0.0686	0.0832	0.0836
40	7	0.0530	0.0501	0.0580	**0.0378**	0.0549	0.0573	0.0614	0.0614
	9	0.0598	0.0585	0.0744	**0.0467**	0.0620	0.0716	0.0802	0.0802
GM 0	7	0.1246	0.1264	0.1218	0.1462	0.1341	0.1154	0.1155	**0.1146**
	9	0.2219	0.2200	0.1971	0.2701	0.2556	0.1951	0.1880	**0.1873**
20	7	0.1323	0.1379	0.1356	0.1626	0.1446	0.1272	0.1298	**0.1288**
	9	0.2257	0.2270	0.2046	0.5204	0.2062	0.2040	0.1983	**0.1968**
40	7	0.1732	0.1807	0.1724	0.2147	0.1762	0.1670	0.1675	**0.1667**
	9	0.2407	0.2489	0.2293	0.2753	0.2450	0.2233	0.2180	**0.2170**
CSF 0	7	0.1632	0.1357	**0.0898**	0.1401	0.1000	0.1309	0.0942	0.0989
	9	0.3023	0.2387	0.1323	0.3356	0.2486	0.2071	**0.1275**	0.1346
20	7	0.1584	0.1414	**0.1006**	0.1476	0.1091	0.1319	0.1061	0.1119
	9	0.3308	0.2737	0.1618	0.7310	0.2543	0.2523	**0.1578**	0.1666
40	7	0.2023	0.1911	**0.1241**	0.2054	0.1489	0.1765	0.1278	0.1367
	9	0.3217	0.2693	0.1595	0.3727	0.2710	0.2339	**0.1567**	0.1666

Table 13 FNR for BrainWeb simulated MRI

INU	Noise	FEC	FCM	IFCM	UPCA	PTFECM	KFEC (Pham et al. 2018)	KIFECM	
								SNF	YNF
WM 0	7	0.0894	0.0885	0.0766	0.1039	0.0894	0.0791	**0.0722**	**0.0722**
	9	0.1659	0.1605	0.1308	0.2046	0.1798	0.1390	**0.1225**	**0.1225**
20	7	0.0962	0.0989	0.0880	0.1184	0.0989	0.0891	**0.0841**	**0.0841**
	9	0.1714	0.1692	0.1395	0.4080	0.1441	0.1495	**0.1331**	**0.1326**
40	7	0.1284	0.1330	0.1193	0.1610	0.1249	0.1209	**0.1155**	**0.1155**
	9	0.1818	0.1847	0.1589	0.2095	0.1768	0.1631	**0.1488**	**0.1488**
GM 0	7	0.1528	0.1417	**0.1280**	0.1309	0.1237	0.1474	0.1334	0.1344
	9	0.2130	0.1877	**0.1591**	0.2023	0.1668	0.1881	0.1641	0.1652
20	7	0.1325	0.1218	**0.1110**	0.1095	0.1060	0.1279	0.1165	0.1167
	9	0.2055	0.1831	**0.1569**	0.3068	0.1860	0.1887	0.1613	0.1630
40	7	0.1458	0.1364	**0.1159**	0.1264	0.1234	0.1394	0.1207	0.1226
	9	0.2062	0.1831	**0.1545**	0.2092	0.1816	0.1862	0.1590	0.1603
CSF 0	7	0.2401	0.2431	0.2574	**0.2336**	0.2486	0.2448	0.2543	0.2499
	9	0.2513	0.2516	0.2530	**0.2159**	0.2271	0.2554	0.2526	0.2465
20	7	0.1996	0.2016	0.2197	**0.1921**	0.2105	0.2061	0.2166	0.2142
	9	0.2234	0.2210	0.2210	0.1642	0.2084	0.2231	0.2210	**0.2139**
40	7	0.1955	0.1976	0.2016	**0.1789**	0.1979	0.2006	0.1986	0.1952
	9	0.2292	0.2329	0.2356	**0.2033**	0.2149	0.2326	0.2356	0.2295

mendable ASA values as shown in the corresponding row in Table 14 as usually there is trade-off between time and ASA.

Results on real brain images

To check the effectiveness of the proposed KIFECM method in comparison with existing methods, experiments have been performed on 2D axial slices of T1-weighted real normal

Table 14 ASA for BrainWeb simulated MRI

INU	Noise	FEC	FCM	IFCM	UPCA	PTFECM	KFEC (Pham et al. 2018)	KIFECM	
								SNF	YNF
0	7	0.8647	0.8688	0.8775	0.8669	0.8743	0.8709	0.8781	**0.8784**
	9	0.8040	0.8159	0.8406	0.7946	0.8180	0.8255	0.8428	**0.8433**
20	7	0.8750	0.8774	0.8839	0.8740	0.8819	0.8792	0.8843	**0.8845**
	9	0.8082	0.8180	0.8420	0.6657	0.8308	0.8250	0.8434	**0.8440**
40	7	0.8552	0.8562	0.8697	0.8492	0.8648	0.8604	**0.8702**	0.8700
	9	0.8021	0.8087	0.8314	0.7916	0.8157	0.8180	0.8346	**0.8350**

Table 15 PC for BrainWeb simulated MRI

INU	Noise	FEC	FCM	IFCM	UPCA	PTFECM	KFEC (Pham et al. 2018)	KIFECM	
								SNF	YNF
0	7	0.9402	0.9055	0.9109	0.8722	0.9383	0.9684	**0.9828**	0.9827
	9	0.9218	0.8903	0.8979	0.8317	0.8891	0.9717	0.9793	**0.9794**
20	7	0.9444	0.9043	0.9094	0.8707	0.9363	0.9773	**0.9826**	**0.9826**
	9	0.9223	0.8899	0.8967	0.7869	0.9090	0.9711	**0.9785**	**0.9785**
40	7	0.9415	0.9003	0.9058	0.8662	0.9318	0.9764	**0.9826**	**0.9826**
	9	0.9284	0.8906	0.8977	0.8458	0.9057	0.9733	**0.9814**	**0.9814**

Table 16 PE for BrainWeb simulated MRI

INU	Noise	FEC	FCM	IFCM	UPCA	PTFECM	KFEC (Pham et al. 2018)	KIFECM	
								SNF	YNF
0	7	0.1452	0.2625	0.2460	0.3093	0.1732	0.0770	**0.0406**	**0.0406**
	9	0.1857	0.2998	0.2779	0.3785	0.2993	0.0677	0.0483	**0.0482**
20	7	0.1354	0.2661	0.2503	0.3126	0.1783	0.0548	**0.0409**	**0.0409**
	9	0.1847	0.3013	0.2809	0.4525	0.2475	0.0690	**0.0501**	0.0502
40	7	0.1413	0.2760	0.2592	0.3200	0.1911	0.0569	**0.0409**	0.0410
	9	0.1708	0.3000	0.2793	0.3521	0.2565	0.0638	**0.0434**	**0.0434**

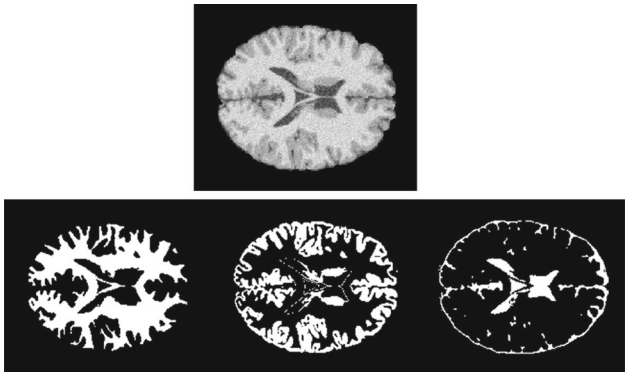


Fig. 5 Original image (INU = 0% and noise level = 7%) and ground truth corresponding to WM, GM and CSF

brain MR images from IBSR database for which ground truth is available. Tables 18, 19, 20 and 21 show the results obtained in terms of DS, JS, FPR and FNR from FEC, FCM, IFCM, UPCA, PTFECM, KFEC (Pham et al. 2018) and the proposed method for GM corresponding to 14 IBSR real brain MRI images on slice 134. Tables 22, 23, 24 and 25 indicate the comparison in terms of DS, JS, FPR and FNR for WM. It can be noted from these tables that the performance of the proposed approach in most of the cases is satisfactory in terms of the performance measures except FPR and FNR for GM and WM. It is worth to mention that many of the images in this dataset having INU due to which the performance in terms of FPR and FNR values for different tissues is not good. Table 26 shows the average segmentation accuracy ASA for 14 cases of IBSR real brain MR images on 134th slice for different methods. It is observed from Table 26 that the proposed KIFECM method with SNF and YNF outperforms the others related method in terms of ASA for different cases. Finally, Table 27 shows the average performance measures over all 14 cases of IBSR real brain MR image for FEC, FCM, IFCM, UPCA, PTFECM, KFEC (Pham et al. 2018) and the proposed method. From Table 27, it can be noted that the proposed KIFECM method with SNF

and YNF achieved superior performance in comparison with FEC, FCM, IFCM, UPCA, PTFECM and KFEC (Pham et al. 2018) methods in terms of all the performance measures.

4.4 Statistical test

A two-way non-parametric statistical test is carried out to investigate the significant difference among the proposed and other segmentation methods using well-known Friedman test for both the publicly available brain MRI datasets. The null hypothesis (H_0) and the alternative hypothesis (H_1) of this test are defined as there is no significant difference in the performance of the proposed and other methods, and the performance of the proposed and other methods is different, respectively, which is stated as:

$$\begin{aligned} H_0 : \mu_{\text{FEC}} &= \mu_{\text{FCM}} = \mu_{\text{IFCM}} = \mu_{\text{UPCA}} \\ &= \mu_{\text{PTFECM}} = \mu_{\text{KFEC[49]}} = \underbrace{\mu_{\text{KIFECM}}}_{\text{SNF}} \\ &= \underbrace{\mu_{\text{KIFECM}}}_{\text{YNF}} \end{aligned} \tag{29}$$

$$\begin{aligned} H_1 : \mu_{\text{FEC}} &\neq \mu_{\text{FCM}} \neq \mu_{\text{IFCM}} \neq \mu_{\text{UPCA}} \\ &\neq \mu_{\text{PTFECM}} \neq \mu_{\text{KFEC[49]}} \neq \underbrace{\mu_{\text{KIFECM}}}_{\text{SNF}} \\ &\neq \underbrace{\mu_{\text{KIFECM}}}_{\text{YNF}} \end{aligned} \tag{30}$$

, respectively, for a given performance measure. In Friedman test, the average rank R_j of jth method for a given M number of images is calculated with respect to a given performance measure as:

$$R_j = \frac{1}{M} \sum_{i=1}^{M} r_i^j \tag{31}$$

where $r_i^j \in \{1, 2, \ldots, t\} (1 \leq i \leq M, 1 \leq j \leq t)$ is rank value for ith image and jth method. Table 28 shows the average Friedman ranking of different methods corresponding to performance measure ASA for 6 BrainWeb brain images and 14 IBSR brain images (Derrac et al. 2011; Friedman 1937). Lowest numerical value of rank for a segmentation method indicates its better performance compared to other methods. It can be noted from Table 28 that the proposed KIFECM

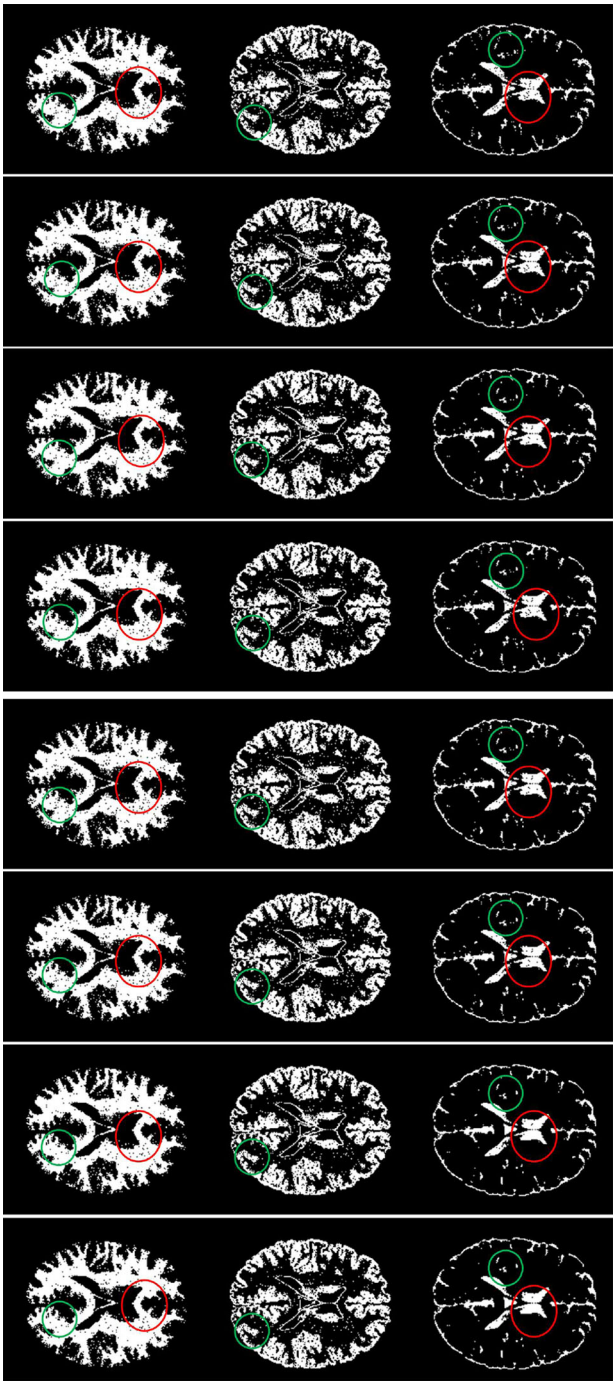


Fig. 6 Qualitative results obtained corresponding to Fig. 5 is given in row-wise fashion. Row 1: FEC, Row 2: FCM, Row 3: IFCM, Row 4: KFEC (Pham et al. 2018), Row 5: UPCA, Row 6: PTFECM, Row 7: KIFECM with SNF and Row 8: KIFECM with YNF

Table 17 Average computation time (in sec. with standard deviation) for different methods on BrainWeb image (INU 40% and noise 7%) of size 181 × 217

FEC	FCM	IFCM	UPCA	PTFECM	KFEC (Pham et al. 2018)	KIFECM	
						SNF	YNF
0.64 (0.14)	1.68 (0.14)	0.94 (0.11)	2.03 (0.11)	2.00 (0.21)	0.72(0.15)	1.55 (0.16)	1.57 (0.14)

Table 18 DS for IBSR dataset for GM

Cases	FEC	FCM	IFCM	UPCA	PTFECM	KFEC (Pham et al. 2018)	KIFECM	
							SNF	YNF
100_23	0.7614	0.8191	0.8011	0.5953	0.4637	0.7968	0.8147	**0.8234**
110_3	0.6126	0.7951	0.7723	0.7603	0.6088	0.7735	0.7857	**0.8144**
111_2	0.6614	0.7793	0.7746	0.7220	0.7015	0.7729	0.7731	**0.7895**
112_2	0.7696	0.7729	0.7744	0.7720	0.7660	0.7642	0.7731	**0.7893**
11_3	0.7700	0.8151	0.8008	0.8005	0.7374	0.7901	**0.8164**	0.8151
12_3	0.7835	0.8129	**0.8164**	0.7754	0.7558	0.8065	0.8145	0.8123
13_3	0.7446	**0.8564**	0.8516	0.8465	0.8265	0.8470	0.8532	0.8510
15_3	0.5872	0.6171	0.6711	0.5822	0.5348	0.6274	**0.7006**	0.6270
16_3	0.6730	0.7226	**0.7472**	0.6705	0.4009	0.7373	0.7369	0.7372
17_3	0.7189	0.7232	0.7424	0.7183	0.6789	0.7376	**0.7580**	0.7410
191_3	0.7878	0.7963	0.7969	0.7878	0.7716	0.7906	0.7976	**0.8146**
1_24	0.7460	0.7481	0.7582	0.6830	0.6552	0.7470	0.7567	**0.7684**
202_3	0.7923	0.8331	0.8139	0.6231	0.7336	0.8063	0.8239	**0.8372**
205_3	0.5004	0.8081	0.7930	0.7980	0.7870	0.7946	**0.8159**	0.8104

Table 19 JS for IBSR dataset for GM

Cases	FEC	FCM	IFCM	UPCA	PTFECM	KFEC (Pham et al. 2018)	KIFECM	
							SNF	YNF
100_23	0.6147	0.6937	0.6682	0.4238	0.3018	0.6622	0.6873	**0.6998**
110_3	0.4415	0.6599	0.6290	0.6133	0.4376	0.6306	0.6471	**0.6869**
111_2	0.4941	0.6384	0.6321	0.5650	0.5402	0.6299	0.6301	**0.6522**
112_2	0.6255	0.6299	0.6318	0.6287	0.6207	0.6183	0.6301	**0.6520**
11_3	0.6260	0.6880	0.6678	0.6674	0.5841	0.6530	**0.6897**	0.6880
12_3	0.6440	0.6848	**0.6898**	0.6332	0.6075	0.6758	0.6870	0.6840
13_3	0.5931	**0.7488**	0.7415	0.7338	0.7043	0.7346	0.7440	0.7406
15_3	0.4156	0.4463	0.5050	0.4106	0.3650	0.4571	**0.5392**	0.4567
16_3	0.5072	0.5657	**0.5964**	0.5044	0.2507	0.5839	0.5834	0.5838
17_3	0.5612	0.5664	0.5903	0.5604	0.5139	0.5843	**0.6103**	0.5886
191_3	0.6499	0.6616	0.6624	0.6499	0.6282	0.6537	0.6633	**0.6872**
1_24	0.5950	0.5975	0.6106	0.5186	0.4872	0.5961	0.6087	**0.6239**
202_3	0.6560	0.7139	0.6861	0.4526	0.5792	0.6755	0.7005	**0.7201**
205_3	0.3337	0.6780	0.6570	0.6639	0.6488	0.6592	**0.6890**	0.6812

method for SNF performs the best. The statistic F_{ID}, suggested by Iman and Davenport (1980), is used:

$$F_{\mathrm{ID}} = \frac{(M-1)\chi_F^2}{M(t-1)-\chi_F^2} \tag{32}$$

which is distributed according to F-distribution with $t-1$ and $(t-1)(M-1)$ degrees of freedom and χ_F^2 is the Friedman's statistic defined as $\frac{12M}{t(t+1)}\left[\sum_j R_j^2 - \frac{t(t+1)^2}{4}\right]$ where t and M are number of methods and number of images used for calculating the average rank, respectively, in our experiments $t = 7$ and $M = 20$. The p value obtained by Iman and Davenport statistic is 1.11E–16 for the performance measure ASA, which endorses the rejection of null hypothesis H_0 at the significance level of 0.05. However, the p value so obtained is not reasonable for comparison with the control method, i.e., the one that emerges with the lowest rank. So adjusted p value (Derrac et al. 2011) is computed which takes into account the error accumulated and provides the correct correlation. This adjusted p values are calculated for different methods with respect to a control method which is the proposed KIFECM method with SNF (lowest rank). For this, a set of post hoc procedures are defined and adjusted p value is computed. The most widely used post hoc method (Derrac et al. 2011) to calculate adjusted p values is Holm procedure. Table 29 shows the adjusted p value for different methods corresponding to ASA which indicates the perfor-

Table 20 FPR for IBSR dataset for GM

Cases	FEC	FCM	IFCM	UPCA	PTFECM	KFEC (Pham et al. 2018)	KIFECM	
							SNF	YNF
100_23	**0.0582**	0.0623	0.0732	0.0307	0.1859	0.0616	0.0757	0.0651
110_3	0.1388	**0.0794**	0.0856	0.0591	0.3086	0.0743	0.0868	0.0871
111_2	0.1479	0.1007	0.0947	0.0760	0.1270	0.0916	**0.0932**	0.1104
112_2	0.0700	0.0700	0.0588	0.0735	0.0878	0.0606	**0.0556**	0.0850
11_3	**0.0600**	0.0834	0.0880	0.0752	0.1092	0.0766	0.0904	0.0834
12_3	**0.0731**	0.0800	0.0751	0.0969	0.1486	0.0766	0.0740	0.0771
13_3	0.1110	0.1048	**0.0977**	0.0982	0.1172	0.0914	0.1051	0.1009
15_3	0.2365	0.2430	0.1957	0.2080	0.2197	0.2242	0.2084	0.2252
16_3	0.3475	0.2702	0.2326	0.1372	0.3087	0.2087	0.2260	0.2252
17_3	0.2037	0.2166	**0.1969**	0.2252	0.2484	0.1994	0.2234	0.2127
191_3	0.0430	0.0499	0.0417	0.0430	0.0535	0.0440	0.0397	0.0689
1_24	0.1019	0.1081	0.0967	0.1094	0.1815	0.0991	**0.1002**	0.1256
202_3	**0.0306**	0.0347	0.0368	0.0143	0.0811	0.0312	0.0431	0.0362
205_3	0.1501	**0.0591**	0.0648	0.0527	0.0773	0.0568	0.0626	0.0597

Table 21 FNR for IBSR dataset for GM

Cases	FEC	FCM	IFCM	UPCA	PTFECM	KFEC (Pham et al. 2018)	KIFECM	
							SNF	YNF
100_23	0.3495	0.2631	0.2829	0.5632	0.6421	0.2969	0.2606	**0.2547**
110_3	0.4972	0.2877	0.3171	0.3505	0.4274	0.3225	0.2968	**0.2533**
111_2	0.4328	0.2974	0.3080	0.3921	0.3912	0.3124	0.3111	**0.2758**
112_2	0.3307	0.3260	0.3310	0.3251	0.3248	0.3442	0.3348	**0.2926**
11_3	0.3364	0.2547	0.2734	0.2824	0.3522	0.2970	**0.2479**	0.2547
12_3	0.3089	0.2604	**0.2584**	0.3054	0.3023	0.2724	0.2621	0.2633
13_3	0.3411	**0.1728**	0.1861	0.1941	0.2131	0.1983	0.1778	0.1846
15_3	0.4861	0.4452	**0.3962**	0.5039	0.5548	0.4404	0.3484	0.4404
16_3	0.3165	0.2814	**0.2649**	0.4264	0.6719	0.2942	0.2847	0.2847
17_3	0.3245	0.3109	0.2934	0.3134	0.3585	0.2991	**0.2534**	0.2863
191_3	0.3222	0.3054	0.3100	0.3222	0.3383	0.3176	0.3104	**0.2654**
1_24	0.3444	0.3379	0.3304	0.4247	0.4244	0.3448	0.3304	**0.2978**
202_3	0.3239	0.2613	0.2886	0.5410	0.3738	0.3034	0.2693	**0.2539**
205_3	0.6162	0.2819	0.3004	0.3011	0.3011	0.3033	**0.2679**	0.2781

mance of the proposed KIFECM method with SNF and YNF in terms of ASA has no significant difference. It can be noted that there is statistical difference in terms of ASA between the proposed method and other methods except IFCM.

5 Conclusion

The segmentation of MRI human brain image is a challenging task due to noise and uncertainty present between the boundaries of soft brain tissues that results in improper segmentation. In this paper, we proposed the kernel intuitionistic fuzzy entropy c-means (KIFECM) algorithm to overcome the problem of the FEC algorithm. The KIFECM algorithm exploits the advantage of intuitionistic fuzzy sets and kernel function successfully and handles the uncertainty in the final localization of feature vector in the term of the pixel value. We have compared the performance of the proposed KIFECM method with existing FEC, FCM, IFCM, UPCA, PTFECM and KFEC (Pham et al. 2018) methods in terms of several performance measures (such as partition coefficient, partition entropy, average segmentation accuracy, dice score, Jaccard score, false positive ratio and false negative ratio) on two synthetic datasets, Iris dataset, simulated and real T1-weighted brain MRI images. From the experimental results on different datasets which contain different types of challenges. For triangle dataset, the proposed KIFECM method outperforms other related methods due to exploiting the kernel

Table 22 DS for IBSR dataset for WM

Cases	FEC	FCM	IFCM	UPCA	PTFECM	KFEC (Pham et al. 2018)	KIFECM	
							SNF	YNF
100_23	0.5813	**0.6674**	0.6072	0.5015	0.4198	0.6270	0.6270	**0.6674**
110_3	0.6087	0.7470	0.7137	0.7395	0.5499	0.7252	0.7252	**0.7616**
111_2	0.7237	0.8017	0.7917	0.7917	0.7727	0.7979	0.7861	**0.8038**
112_2	0.7904	0.7904	0.7862	0.7954	0.7877	0.7883	0.7835	**0.7966**
11_3	0.7443	**0.7505**	0.7246	0.7505	0.7024	0.7246	0.7443	**0.7505**
12_3	0.8446	**0.8494**	0.8358	0.8510	0.8192	0.8446	0.8298	0.8446
13_3	**0.8764**	0.8744	0.8690	0.8744	0.8559	0.8690	0.8690	0.8707
15_3	0.6405	0.6405	**0.6856**	0.6586	0.6586	0.6586	**0.6856**	0.6586
16_3	0.6268	0.5856	0.6351	0.6448	0.5548	0.6409	**0.6448**	0.6278
17_3	0.7058	0.7016	0.7177	0.6920	**0.7181**	0.7148	0.7148	0.7119
191_3	0.8151	0.8193	0.8084	0.8151	0.7995	0.8151	0.8043	**0.8280**
1_24	0.8064	0.8085	0.7918	0.8124	0.7756	0.8030	0.7848	**0.8124**
202_3	0.6859	**0.7363**	0.6859	0.5855	0.6486	0.7027	0.6859	**0.7363**
205_3	0.5103	**0.7419**	0.7108	0.7419	0.7307	0.7241	**0.7419**	**0.7419**

Table 23 JS for IBSR dataset for WM

Cases	FEC	FCM	IFCM	UPCA	PTFECM	KFEC (Pham et al. 2018)	KIFECM	
							SNF	YNF
100_23	0.4098	0.5008	0.4359	0.3347	0.2656	0.4566	0.4566	**0.5008**
110_3	0.4375	0.5962	0.5549	0.5866	0.3792	0.5689	0.5689	**0.6150**
111_2	0.5670	0.6691	0.6552	0.6552	0.6296	0.6638	0.6475	**0.6720**
112_2	0.6535	0.6535	0.6477	0.6603	0.6497	0.6506	0.6440	**0.6620**
11_3	0.5928	**0.6007**	0.5682	0.6007	0.5413	0.5682	0.5928	**0.6007**
12_3	0.7311	**0.7382**	0.7179	0.7406	0.6938	0.7311	0.7090	0.7311
13_3	**0.7800**	0.7769	0.7684	0.7769	0.7480	0.7684	0.7684	0.7710
15_3	0.4711	0.4711	0.5216	0.4910	0.4910	0.4910	0.5216	0.4910
16_3	0.4564	0.4140	0.4653	0.4758	0.3839	0.4716	**0.4758**	0.4575
17_3	0.5453	0.5404	**0.5596**	0.5290	0.5602	0.5562	0.5562	0.5526
191_3	0.6880	0.6939	0.6784	0.6880	0.6660	0.6880	0.6726	**0.7065**
1_24	0.6756	0.6785	0.6553	0.6841	0.6334	0.6708	0.6459	**0.6841**
202_3	0.5220	**0.5827**	0.5220	0.4140	0.4799	0.5417	0.5220	**0.5827**
205_3	0.3425	**0.5897**	0.5514	0.5897	0.5757	0.5676	**0.5897**	**0.5897**

Table 24 FPR for IBSR dataset for WM

Cases	FEC	FCM	IFCM	UPCA	PTFECM	KFEC (Pham et al. 2018)	KIFECM	
							SNF	YNF
100_23	1.2631	0.8110	1.1206	1.8183	2.5567	1.0116	1.0116	**0.8110**
110_3	1.2229	0.5405	0.6996	0.5796	1.5700	0.6468	0.6468	**0.4767**
111_2	0.7041	0.3115	0.3770	0.3770	0.4330	0.3493	0.4143	**0.2893**
112_2	0.3715	0.3715	0.4257	0.3434	0.3654	0.4011	0.4456	**0.3168**
11_3	0.5330	0.4860	0.6278	0.4860	0.7086	0.6278	0.5330	**0.4860**
12_3	0.2259	0.2067	0.2643	**0.1527**	0.2305	0.2259	0.2949	0.2259
13_3	**0.1082**	0.1665	0.2088	0.1665	0.2293	0.2088	0.2088	0.1927
15_3	**0.3108**	0.3108	0.3978	0.3425	0.3425	0.3425	0.3978	0.3425
16_3	1.0664	**0.5318**	0.7071	0.8588	1.4739	0.7507	0.8588	0.6512

Table 24 continued

Cases	FEC	FCM	IFCM	UPCA	PTFECM	KFEC (Pham et al. 2018)	KIFECM	
							SNF	YNF
17_3	0.4394	**0.4073**	0.5156	0.3928	0.5385	0.4729	0.4729	0.4561
191_3	0.3508	0.3224	0.3797	0.3508	0.4053	0.3508	0.4031	**0.2539**
1_24	0.3453	0.3215	0.4315	0.2768	0.3790	0.3696	0.4619	**0.2768**
202_3	0.7702	0.5621	0.7702	1.2870	0.9379	0.7022	0.7702	**0.5621**
205_3	1.8039	0.5552	0.6851	0.5552	0.5930	0.6289	**0.5552**	**0.5552**

Table 25 FNR for IBSR dataset for WM

Cases	FEC	FCM	IFCM	UPCA	PTFECM	KFEC (Pham et al. 2018)	KIFECM	
							SNF	YNF
100_23	0.0727	0.0930	0.0756	0.0567	**0.0552**	0.0814	0.0814	0.0930
110_3	0.0274	0.0816	0.0569	0.0734	**0.0254**	0.0631	0.0631	0.0919
111_2	**0.0338**	0.1225	0.0978	0.0978	0.0978	0.1043	0.0842	0.1336
112_2	0.1037	0.1037	0.0766	0.1129	0.1129	0.0884	**0.0690**	0.1283
11_3	0.0913	0.1074	**0.0751**	0.1074	**0.0751**	0.0751	0.0913	0.1074
12_3	0.1038	0.1093	0.0924	0.1463	0.1463	0.1038	**0.0818**	0.1038
13_3	0.1356	0.0937	**0.0712**	0.0937	0.0805	0.0712	**0.0712**	0.0805
15_3	0.3825	0.3825	0.2709	0.3408	0.3408	0.3408	**0.2709**	0.3408
16_3	0.0569	0.3659	0.2057	0.1156	**0.0502**	0.1744	0.1156	0.2445
17_3	0.2151	0.2395	**0.1518**	0.2632	0.1381	0.1808	0.1808	0.1953
191_3	0.0707	0.0824	0.0640	0.0707	0.0640	0.0707	**0.0562**	0.1141
1_24	0.0912	0.1033	0.0619	0.1265	0.1265	0.0812	**0.0558**	0.1265
202_3	0.0759	0.0897	0.0759	0.0533	**0.0700**	0.0779	0.0759	0.0897
205_3	0.0396	0.0829	**0.0709**	0.0829	0.0829	0.0755	0.0829	0.0829

Table 26 ASA for IBSR dataset

Cases	FEC	FCM	IFCM	UPCA	PTFECM	KFEC (Pham et al. 2018)	KIFECM	
							SNF	YNF
100_23	0.6992	0.7666	0.7535	0.5272	0.4627	0.7409	**0.7707**	**0.7735**
110_3	0.6385	0.7713	0.7577	0.7298	0.6890	0.7521	0.7703	**0.7927**
111_2	0.7074	0.7614	0.7628	0.7150	0.7163	0.7586	**0.7648**	**0.7697**
112_2	0.7548	0.7576	0.7636	0.7553	0.7555	0.7523	**0.7640**	**0.7690**
11_3	0.7318	0.7859	0.7811	0.7662	0.7250	0.7641	**0.7949**	0.7859
12_3	0.7689	0.7950	0.8012	0.7562	0.7588	0.7899	**0.8021**	0.7956
13_3	0.7453	0.8461	0.8451	0.8365	0.8317	0.8395	**0.8485**	0.8428
15_3	0.5584	0.5834	0.6534	0.5627	0.5316	0.6005	**0.6818**	0.6003
16_3	0.7616	0.6924	0.7524	0.6673	0.5167	0.7415	**0.7660**	0.7269
17_3	0.7096	0.7111	0.7498	0.7021	0.7113	0.7369	**0.7676**	0.7411
191_3	0.7701	0.7763	0.7802	0.7701	0.7625	0.7730	0.7829	**0.7897**
1_24	0.7514	0.7510	0.7656	0.6935	0.6955	0.7541	0.7646	**0.7658**
202_3	0.7336	0.7784	0.7607	0.5732	0.6968	0.7489	0.7755	**0.7841**
205_3	0.5286	0.7667	0.7558	0.7530	0.7530	0.7528	**0.7768**	**0.7695**

properties and correctly partition the boundary data points. Again the results on noisy gray image datasets show the proposed KIFECM method is robust to noise when compared to other related methods due to the advantage of IFSs and kernel distance measures. Similarly, experimental results on other dataset show the superior performance of the proposed KIFECM method over FEC, FCM, IFCM, UPCA, PTFECM and KFEC (Pham et al. 2018) methods. The Friedman sta-

Table 27 Average performance measures over 14 cases of IBSR real brain MRI (slice no. 134)

		FEC	FCM	IFCM	UPCA	PTFECM	KFEC (Pham et al. 2018)	KIFECM	
								SNF	YNF
DS	WM	0.7114	0.7510	0.7402	0.7325	0.6995	0.7454	0.7448	**0.7580**
	GM	0.7078	0.7785	0.7796	0.7239	0.6730	0.7708	0.7872	**0.7879**
JS	WM	0.5623	0.6075	0.5930	0.5876	0.5498	0.5996	0.5979	**0.6155**
	GM	0.5541	0.6409	0.6406	0.5732	0.5192	0.6296	0.6507	**0.6532**
FPR	WM	0.6797	0.4218	0.5436	0.5705	0.7688	0.5064	0.5339	**0.4212**
	GM	0.1266	0.1116	0.1027	**0.0928**	0.1610	0.0997	0.1060	0.1116
FNR	WM	0.1072	0.1470	0.1033	0.1244	0.1047	0.1135	**0.0986**	0.1380
	GM	0.3807	0.2919	0.2958	0.3747	0.4054	0.3105	0.2825	**0.2775**
Accuracy	Overall cases	0.7042	0.7531	0.7631	0.7006	0.6862	0.7504	**0.7736**	0.7647
PC	Overall cases	0.8249	0.8591	0.8625	0.7275	0.8362	0.9528	**0.9691**	0.9376
PE	Overall cases	0.3876	0.3813	0.3691	0.2777	0.4356	0.1124	**0.0712**	0.1413

Table 28 Friedman ranking of different methods

	FEC	FCM	IFCM	UPCA	PTFECM	KFEC (Pham et al. 2018)	KIFECM	
							SNF	YNF
ASA	6.73	4.48	3.05	7.05	6.12	4.95	**1.75**	1.86

Table 29 Adjusted p values (Friedman) corresponding to performance measure ASA

	PTFECM	UPCA	KFEC (Pham et al. 2018)	FEC	FCM	IFCM	KIFECM with YNF
Unadjusted p value	1.62E–08	7.79E–12	3.61E–05	1.34E–10	4.35E–04	0.0932	0.8718
Holm value	8.11E–08	5.46E–11	1.44E–04	8.03E–10	0.0013	0.1866	0.8718

tistical test is performed for both the publicly available brain MRI datasets to investigate the significant difference among the proposed KIFECM method and other related methods. The statistical test endorses the performance of proposed method over other related method. This work does not consider spatial information to overcome the problem of noise and intensity inhomogeneity. The proper incorporation of spatial information in the proposed approach to address the problem of noise and intensity inhomogeneity may be an interesting future goal.

Acknowledgements The authors would like to thank Department of Science and Technology (PURSE), New Delhi, and Council of Scientific & Industrial Research, New Delhi, for the financial support. The authors are also thankful to the anonymous reviewers for their constructive suggestions.

Compliance with ethical standards

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

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

A Derivation for partition matrix and cluster centroid

The Lagrangian for the objective function (23) of the proposed KIFECM method can be given as:

$$\begin{aligned} L = & \sum_{i=1}^{c}\sum_{j=1}^{N}\mu_{ij}(1-K(\mathbf{x}_j^{\mathrm{IFS}},\mathbf{v}_i^{\mathrm{IFS}})) \\ & + \frac{\sigma^2}{c}\sum_{i=1}^{c}\sum_{j=1}^{N}(\mu_{ij}\log\mu_{ij}-\mu_{ij}) \\ & + \lambda\sum_{i=1}^{c}\sum_{j=1}^{N}\mu_{ij}\log\mu_{ij} \\ & + \sum_{i=1}^{N}Y_j\left(1-\sum_{i=1}^{c}\mu_{ij}\right) \end{aligned} \tag{33}$$

where Y_j is Lagrange multiplier $\forall j \in \{1,2,\ldots,N\}$. The solution can be obtained after differentiating above with respect to μ_{ij}, $\mu_V(v_i)$, $\nu_V(v_i)$ and $\pi_V(v_i)$.

$$\begin{aligned} \frac{\partial L}{\partial \mu_{ij}} &= (1-K(\mathbf{x}_j^{\mathrm{IFS}},\mathbf{v}_i^{\mathrm{IFS}})) + \frac{\sigma^2}{c}\log\mu_{ij} \\ &\quad + \lambda(1+\log\mu_{ij}) - Y_j = 0 \\ &\Rightarrow c(1-K(\mathbf{x}_j^{\mathrm{IFS}},\mathbf{v}_i^{\mathrm{IFS}})) + (\sigma^2+c\lambda)\log\mu_{ij} \end{aligned}$$

$$+c\lambda - cY_j = 0$$
$$\Rightarrow (\sigma^2 + c\lambda)\log\mu_{ij} = cY_j - c(1 - K(\mathbf{x}_j^{\mathrm{IFS}}, \mathbf{v}_i^{\mathrm{IFS}})) - c\lambda$$
$$\Rightarrow (\sigma^2 + c\lambda)\log\mu_{ij} = cY_j - c[(1 - K(\mathbf{x}_j^{\mathrm{IFS}}, \mathbf{v}_i^{\mathrm{IFS}})) + \lambda]$$
$$\Rightarrow \log\mu_{ij} = \frac{cY_j}{(\sigma^2 + c\lambda)} - \frac{c[(1 - K(\mathbf{x}_j^{\mathrm{IFS}}, \mathbf{v}_i^{\mathrm{IFS}})) + \lambda]}{(\sigma^2 + c\lambda)}$$
$$\Rightarrow \mu_{ij} = \exp\left(\frac{cY_j}{(\sigma^2 + c\lambda)}\right) \exp\left(-\frac{c[(1 - K(\mathbf{x}_j^{\mathrm{IFS}}, \mathbf{v}_i^{\mathrm{IFS}})) + \lambda]}{(\sigma^2 + c\lambda)}\right)$$

Elimination of Y_j with the constraint will leads to (34)

$$\Rightarrow \mu_{ij} = \frac{\exp\left(-\frac{c[(1-K(\mathbf{x}_j^{\mathrm{IFS}}, \mathbf{v}_i^{\mathrm{IFS}}))+\lambda]}{(\sigma^2+c\lambda)}\right)}{\sum_{k=1}^{c}\exp\left(-\frac{c[(1-K(\mathbf{x}_j^{\mathrm{IFS}}, \mathbf{v}_k^{\mathrm{IFS}}))+\lambda]}{(\sigma^2+c\lambda)}\right)}$$
$$\Rightarrow \mu_{ij} = \frac{\exp\left(-\frac{[(1-K(\mathbf{x}_j^{\mathrm{IFS}}, \mathbf{v}_i^{\mathrm{IFS}}))+\lambda]}{\left(\frac{\sigma^2}{c}+\lambda\right)}\right)}{\sum_{k=1}^{c}\exp\left(-\frac{[(1-K(\mathbf{x}_j^{\mathrm{IFS}}, \mathbf{v}_k^{\mathrm{IFS}}))+\lambda]}{\left(\frac{\sigma^2}{c}+\lambda\right)}\right)}$$
$$\Rightarrow \mu_{ij} = \frac{\exp(-[(1 - K(\mathbf{x}_j^{\mathrm{IFS}}, \mathbf{v}_i^{\mathrm{IFS}})) + \lambda])^{\frac{1}{\left(\frac{\sigma^2}{c}+\lambda\right)}}}{\sum_{k=1}^{c}\exp(-[(1 - K(\mathbf{x}_j^{\mathrm{IFS}}, \mathbf{v}_k^{\mathrm{IFS}})) + \lambda])^{\frac{1}{\left(\frac{\sigma^2}{c}+\lambda\right)}}}$$
$$\Rightarrow \mu_{ij} = \frac{1}{\sum_{k=1}^{c}\frac{\exp([(1-K(\mathbf{x}_j^{\mathrm{IFS}}, \mathbf{v}_i^{\mathrm{IFS}}))+\lambda])^{\frac{1}{\left(\frac{\sigma^2}{c}+\lambda\right)}}}{\exp([(1-K(\mathbf{x}_j^{\mathrm{IFS}}, \mathbf{v}_k^{\mathrm{IFS}}))+\lambda])^{\frac{1}{\left(\frac{\sigma^2}{c}+\lambda\right)}}}}$$
$$\Rightarrow \mu_{ij} = \left\{\sum_{k=1}^{c}\left(\frac{\exp(1 - K(\mathbf{x}_j^{\mathrm{IFS}}, \mathbf{v}_i^{\mathrm{IFS}}) + \lambda)}{\exp(1 - K(\mathbf{x}_j^{\mathrm{IFS}}, \mathbf{v}_k^{\mathrm{IFS}}) + \lambda)}\right)^{\frac{1}{\frac{\sigma^2}{c}+\lambda}}\right\}^{-1} \quad (35)$$

$$\Rightarrow \mu_{ij} = \left\{\sum_{k=1}^{c}\left(\frac{\exp(1 - K(\mathbf{x}_j^{\mathrm{IFS}}, \mathbf{v}_i^{\mathrm{IFS}}))}{\exp(1 - K(\mathbf{x}_j^{\mathrm{IFS}}, \mathbf{v}_k^{\mathrm{IFS}}))}\right)^{\frac{1}{\frac{\sigma^2}{c}+\lambda}}\right\}^{-1} \quad (36)$$

$$\frac{\partial L}{\partial \mu_V(v_i)} = \sum_{j=1}^{N}\mu_{ij}K(\mathbf{x}_j^{\mathrm{IFS}}, \mathbf{v}_i^{\mathrm{IFS}})(\mu_X(x_j) - \mu_V(v_i)) = 0$$
$$\Rightarrow \sum_{j=1}^{N}\mu_{ij}K(\mathbf{x}_j^{\mathrm{IFS}}, \mathbf{v}_i^{\mathrm{IFS}})\mu_X(x_j) - \sum_{j=1}^{N}\mu_{ij}K(\mathbf{x}_j^{\mathrm{IFS}}, \mathbf{v}_i^{\mathrm{IFS}})\mu_V(v_i) = 0 \quad (37)$$
$$\Rightarrow \mu_V(v_i) = \frac{\sum_{j=1}^{N}\mu_{ij}(K(\mathbf{x}_j^{\mathrm{IFS}}, \mathbf{v}_i^{\mathrm{IFS}})\mu_X(x_j))}{\sum_{j=1}^{N}\mu_{ij}(K(\mathbf{x}_j^{\mathrm{IFS}}, \mathbf{v}_i^{\mathrm{IFS}}))} \quad (38)$$

$$\frac{\partial L}{\partial \nu_V(v_i)} = \sum_{j=1}^{N}\mu_{ij}K(\mathbf{x}_j^{\mathrm{IFS}}, \mathbf{v}_i^{\mathrm{IFS}})(\nu_X(x_j) - \nu_V(v_i)) = 0$$
$$\Rightarrow \sum_{j=1}^{N}\mu_{ij}K(\mathbf{x}_j^{\mathrm{IFS}}, \mathbf{v}_i^{\mathrm{IFS}})\nu_X(x_j) - \sum_{j=1}^{N}\mu_{ij}K(\mathbf{x}_j^{\mathrm{IFS}}, \mathbf{v}_i^{\mathrm{IFS}})\nu_V(v_i) = 0 \quad (39)$$
$$\Rightarrow \nu_V(v_i) = \frac{\sum_{j=1}^{N}\mu_{ij}(K(\mathbf{x}_j^{\mathrm{IFS}}, \mathbf{v}_i^{\mathrm{IFS}})\nu_X(x_j))}{\sum_{j=1}^{N}\mu_{ij}(K(\mathbf{x}_j^{\mathrm{IFS}}, \mathbf{v}_i^{\mathrm{IFS}}))} \quad (40)$$

$$\frac{\partial L}{\partial \pi_V(v_i)} = \sum_{j=1}^{N}\mu_{ij}K(\mathbf{x}_j^{\mathrm{IFS}}, \mathbf{v}_i^{\mathrm{IFS}})(\pi_X(x_j) - \pi_V(v_i)) = 0$$
$$\Rightarrow \sum_{j=1}^{N}\mu_{ij}K(\mathbf{x}_j^{\mathrm{IFS}}, \mathbf{v}_i^{\mathrm{IFS}})\pi_X(x_j) - \sum_{j=1}^{N}\mu_{ij}K(\mathbf{x}_j^{\mathrm{IFS}}, \mathbf{v}_i^{\mathrm{IFS}})\pi_V(v_i) = 0 \quad (41)$$
$$\Rightarrow \pi_V(v_i) = \frac{\sum_{j=1}^{N}\mu_{ij}(K(\mathbf{x}_j^{\mathrm{IFS}}, \mathbf{v}_i^{\mathrm{IFS}})\pi_X(x_j))}{\sum_{j=1}^{N}\mu_{ij}(K(\mathbf{x}_j^{\mathrm{IFS}}, \mathbf{v}_i^{\mathrm{IFS}}))} \quad (42)$$

References

Alipour S, Shanbehzadeh J (2014) Fast automatic medical image segmentation based on spatial kernel fuzzy c-means on level set method. Mach Vis Appl 25(6):1469–1488

Anbeek P, Vincken KL, van Osch MJ, Bisschops RH, van der Grond J (2004) Automatic segmentation of different-sized white matter lesions by voxel probability estimation. Med Image Anal 8(3):205–215

Atanassov KT (1986) Intuitionistic fuzzy sets. Fuzzy Sets Syst 20(1):87–96

Atanassov KT (2003) Intuitionistic fuzzy sets: past, present and future. In: EUSFLAT conference, pp 12–19

Balafar M (2014) Fuzzy c-mean based brain mri segmentation algorithms. Artif Intell Rev 41(3):441–449

Balafar MA, Ramli AR, Saripan MI, Mashohor S (2010) Review of brain mri image segmentation methods. Artif Intell Rev 33(3):261–274

Benaichouche A, Oulhadj H, Siarry P (2013) Improved spatial fuzzy c-means clustering for image segmentation using pso initialization, mahalanobis distance and post-segmentation correction. Dig Signal Process 23(5):1390–1400

Berry MW, Castellanos M (2004) Survey of text mining. Comput Rev 45(9):548

Bezdek JC (1981) Objective function clustering. In: Pattern recognition with fuzzy objective function algorithms. Springer, pp 43–93

Bezdek JC, Keller JM, Krishnapuram R, Kuncheva LI, Pal NR (1999) Will the real iris data please stand up? IEEE Trans Fuzzy Syst 7(3):368–369

Bustince H, Kacprzyk J, Mohedano V (2000) Intuitionistic fuzzy generators application to intuitionistic fuzzy complementation. Fuzzy Sets Syst 114(3):485–504

Chaira T (2011) A novel intuitionistic fuzzy c means clustering algorithm and its application to medical images. Appl Soft Comput 11(2):1711–1717

Chen S, Zhang D (2004) Robust image segmentation using fcm with spatial constraints based on new kernel-induced distance measure. IEEE Trans Syst Man Cybern Part B Cybern 34(4):1907–1916

Chen X, Nguyen BP, Chui CK, Ong SH (2016) Automated brain tumor segmentation using kernel dictionary learning and superpixel-level features. In: 2016 IEEE international conference on systems, man, and cybernetics (SMC). IEEE, pp 002547–002552

Chuang KS, Tzeng HL, Chen S, Wu J, Chen TJ (2006) Fuzzy c-means clustering with spatial information for image segmentation. Comput Med Imaging Graph 30(1):9–15

Cocosco CA, Kollokian V, Kwan RKS, Pike GB, Evans AC (1997) Brainweb: online interface to a 3d mri simulated brain database. In: NeuroImage. Citeseer

Cover TM (1965) Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. IEEE Trans Electron Comput 3:326–334

Derrac J, García S, Molina D, Herrera F (2011) A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. Swarm Evol Comput 1(1):3–18

Fisher RA (1936) The use of multiple measurements in taxonomic problems. Ann Hum Genet 7(2):179–188

Friedman M (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J Am Stat Assoc 32(200):675–701

Graves D, Pedrycz W (2010) Kernel-based fuzzy clustering and fuzzy clustering: a comparative experimental study. Fuzzy Sets Syst 161(4):522–543

Hai-Jun F, Xiao-Hong W, Han-Ping M, Bin W (2011) Fuzzy entropy clustering using possibilistic approach. Procedia Eng 15:1993–1997

Held K, Kops ER, Krause BJ, Wells WM, Kikinis R, Muller-Gartner HW (1997) Markov random field segmentation of brain mr images. IEEE Trans Med Imaging 16(6):878–886

Hoffman EJ, Huang SC, Phelps ME (1979) Quantitation in positron emission computed tomography: 1.Effect of object size. J Comput Assist Tomogr 3(3):299–308

Huang CW, Lin KP, Wu MC, Hung KC, Liu GS, Jen CH (2015) Intuitionistic fuzzy c-means clustering algorithm with neighborhood attraction in segmenting medical image. Soft Comput 19(2):459–470

Iakovidis D, Pelekis N, Kotsifakos E, Kopanakis I (2008) Intuitionistic fuzzy clustering with applications in computer vision. In: Advanced concepts for intelligent vision systems. Springer, pp 764–774

Iman RL, Davenport JM (1980) Approximations of the critical region of the fbietkan statistic. Commun Stat Theory Methods 9(6):571–595

Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. ACM Comput Surv CSUR 31(3):264–323

Ji ZX, Sun QS, Xia DS (2014) A framework with modified fast fcm for brain mr images segmentation (retraction of vol 44, pg 999, 2011). Pattern Recognit 47(12):3979–3979

Kahali S, Sing JK, Saha PK (2018) A new fuzzy clustering algorithm for brain mr image segmentation using gaussian probabilistic and entropy-based likelihood measures. In: 2018 international conference on communication, computing and internet of things (IC3IoT). pp 54–59. https://doi.org/10.1109/IC3IoT.2018.8668139

Kannan S, Devi R, Ramathilagam S, Takezawa K (2013) Effective fcm noise clustering algorithms in medical images. Comput Biol Med 43(2):73–83

Krinidis S, Chatzis V (2010) A robust fuzzy local information c-means clustering algorithm. IEEE Trans Image Process 19(5):1328–1337

Krishnapuram R, Keller JM (1993) A possibilistic approach to clustering. IEEE Trans Fuzzy Syst 1(2):98–110

Kumar D, Verma H, Mehra A, Agrawal RK (2018) A modified intuitionistic fuzzy c-means clustering approach to segment human brain mri image. Multimed Tools Appl. https://doi.org/10.1007/s11042-018-5954-0

Li RP, Mukaidono M (1999) Gaussian clustering method based on maximum-fuzzy-entropy interpretation. Fuzzy Sets Syst 102(2):253–258

Li C, Huang R, Ding Z, Gatenby JC, Metaxas DN, Gore JC (2011) A level set method for image segmentation in the presence of intensity inhomogeneities with application to mri. IEEE Trans Image Process 20(7):2007–2016

Li C, Zhao H, Xu Z (2018) Kernel c-means clustering algorithms for hesitant fuzzy information in decision making. Int J Fuzzy Syst 20(1):141–154

Lin KP (2013) A novel evolutionary kernel intuitionistic fuzzy *c*-means clustering algorithm. IEEE Trans Fuzzy Syst 22(5):1074–1087

Lloyd S (1982) Least squares quantization in pcm. IEEE Trans Inf Theory 28(2):129–137

Mercer J (1909) Xvi. functions of positive and negative type, and their connection the theory of integral equations. Philos Trans R Soc Lond A 209(441–458):415–446

Muller KR, Mika S, Ratsch G, Tsuda K, Scholkopf B (2001) An introduction to kernel-based learning algorithms. IEEE Trans Neural Netw 12(2):181–201

Murofushi T, Sugeno M (2000) Fuzzy measures and fuzzy integrals. In: Grabisch M, Murofushi T, Sugeno M (eds) Fuzzy measures and integrals. Theory and applications. Heidelberg, Physica-Verlag, pp 3–41

Nayak J, Naik B, Behera H (2015) Fuzzy c-means (fcm) clustering algorithm: a decade review from 2000 to 2014. In: Computational intelligence in data mining-volume 2. Springer, pp 133–149

Olabarriaga SD, Smeulders AW (2001) Interaction in the segmentation of medical images: a survey. Med Image Anal 5(2):127–142

Pal NR, Pal SK (1992) Higher order fuzzy entropy and hybrid entropy of a set. Inf Sci 61(3):211–231

Pal NR, Pal K, Keller JM, Bezdek JC (2005) A possibilistic fuzzy c-means clustering algorithm. IEEE Trans Fuzzy Syst 13(4):517–530

Pelekis N, Iakovidis DK, Kotsifakos EE, Kopanakis I (2008) Fuzzy clustering of intuitionistic fuzzy data. Int J Bus Intell Data Min 3(1):45–65

Pham DL, Xu C, Prince JL (2000) Current methods in medical image segmentation. Ann Rev Biomed Eng 2(1):315–337

Pham TX, Siarry P, Oulhadj H (2018) Integrating fuzzy entropy clustering with an improved pso for mri brain image segmentation. Appl Soft Comput 65:230–242

Qiu C, Xiao J, Yu L, Han L, Iqbal MN (2013) A modified interval type-2 fuzzy c-means algorithm with application in mr image segmentation. Pattern Recognit Lett 34(12):1329–1338

Reddick WE, Glass JO, Cook EN, Elkin TD, Deaton RJ (1997) Automated segmentation and classification of multispectral magnetic resonance images of brain using artificial neural networks. IEEE Trans Med Imaging 16(6):911–918

Rohlfing T, Brandt R, Menzel R, Maurer CR (2004) Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. NeuroImage 21(4):1428–1442

Rui Y, Huang TS, Chang SF (1999) Image retrieval: current techniques, promising directions, and open issues. J Vis Commun Image Represent 10(1):39–62

Sato M, Lakare S, Wan M, Kaufman A, Nakajima M (2000) A gradient magnitude based region growing algorithm for accurate segmentation. In: 2000 International Conference on image processing, 2000. Proceedings, vol 3. IEEE, pp 448–451

Schaaf T, Kemp T (1997) Confidence measures for spontaneous speech recognition. In: International conference on acoustics, speech, and signal processing, 1997. ICASSP-97, IEEE, vol 2. IEEE, pp 875–878

Szmidt E, Kacprzyk J (2000) Distances between intuitionistic fuzzy sets. Fuzzy Sets Syst 114(3):505–518

Tran D, Wagner M (2000) Fuzzy entropy clustering. In: The ninth IEEE international conference on fuzzy systems, 2000. FUZZ IEEE 2000, vol 1. IEEE, pp 152–157

Verma H, Agrawal R (2015) Possibilistic intuitionistic fuzzy c-means clustering algorithm for mri brain image segmentation. Int J Artif Intell Tools 24(05):1550,016

Verma H, Agrawal RK, Kumar N (2014) Improved fuzzy entropy clustering algorithm for mri brain image segmentation. Int J Imaging Syst Technol 24(4):277–283

Verma H, Agrawal R, Sharan A (2016) An improved intuitionistic fuzzy c-means clustering algorithm incorporating local information for brain image segmentation. Appl Soft Comput 46:543–557

Vlachos IK, Sergiadis GD (2007) The role of entropy in intuitionistic fuzzy contrast enhancement. In: International fuzzy systems association world congress. Springer, pp 104–113

Vovk U, Pernus F, Likar B (2007) A review of methods for correction of intensity inhomogeneity in mri. IEEE Trans Med Imaging 26(3):405–421

Wang L, Chen Y, Pan X, Hong X, Xia D (2010) Level set segmentation of brain magnetic resonance images based on local gaussian distribution fitting energy. J Neurosci Methods 188(2):316–325

Wang Z, Song Q, Soh YC, Sim K (2013) An adaptive spatial information-theoretic fuzzy clustering algorithm for image segmentation. Computer Vis Image Underst 117(10):1412–1420

Xu Z, Wu J (2010) Intuitionistic fuzzy c-means clustering algorithms. J Syst Eng Electron 21(4):580–590

Yager RR (1979) On the measure of fuzziness and negation part i: membership in the unit interval. Int J General Syst 5(4):221–229

Yager RR (1980) On the measure of fuzziness and negation. ii. Lattices. Inf Control 44(3):236–260

Yager RR (2000) On the entropy of fuzzy measures. IEEE Trans Fuzzy Syst 8(4):453–461

Yang MS, Wu KL (2006) Unsupervised possibilistic clustering. Pattern Recognit 39(1):5–21

Zadeh LA (1965) Fuzzy sets. Inf Control 8(3):338–353

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

Lithium-ion Battery Safety Tests monitored by LabVIEW and Comsol software for electrical and mechanical characteristics.

Snigdha Sharma[a, b], M M Tripathi[a], Amrish K Panwar[b]

[a]Department of Electrical Engineering, Delhi Technological University, New Delhi, India
[b]Department of Applied Physics, Delhi Technological University, New Delhi, India

This paper presents a unique cheaper battery safety test bench for electrical and mechanical performance evaluation of lithium-ion batteries where mechanical performance is tested using software and electrical characteristic is tested using hardware. In this work, simulation of mechanical performance of lithium-ion battery such as thermal modelling test at various temperature has been carried out using Comsol software, where battery voltage and temperature are monitored as a control input during a drive cycle and state of charge of the battery on the positive (65-80%) and a negative electrode (14-29%) is monitored and triggered as a control output. A hardware test bench has been developed to execute various electrical tests for evaluating battery performance during a discharge cycle. These electrical tests have been carried out on the LabVIEW software to obtain real-time data and display the battery related information such as voltage and current curves. Further, a battery safety test like external short circuit test has been performed and also for improving the battery efficiency capacitors are connected parallelly with varied load condition. Overall, in this exploration, a newly evolved, accurate, affordable battery testing and monitoring system has been developed.

Keywords: LabVIEW Software, Comsol Software, Battery Safety test, Electrical Test, Mechanical Test

1. INTRODUCTION

From the last few decades, the lithium-ion battery has gained so much attention due to its miscellaneous advantage in the energy storage area because of high capacity, high energy or high power, long cyclability, low memory effect, and low environmental effect[1]. In general, the batteries in today's world are not only for portable electronics but it also moves towards a large scale electricity storage on smart grids, the automotive industry in terms of electric vehicles (EVs)[2], aerospace engineering, communication, back-up, and many more[3]. Hence, battery technology usage increases rapidly in a wide area for future prospects[2]. So, we focused on lithium-ion batteries but we all know that these are very dangerous and complicated for experiment purpose[4]. So for making these batteries more reliable and safe, the battery test is performed for determining the battery characteristics[5].

In this paper, we present two types of battery characterization techniques i.e. mechanical and electrical characterization. Mechanical characterization techniques have obtained through Multiphysics Software. In this simulation, experimental data was

used to design a Multiphysics based theoretical model for optimizing the battery drive cycle performances[6]. This model setup includes some physical, thermal and electrical properties and then optimized the mechanical results such as polarization test, thermal test (local temperature), drive current and voltage of the battery during the drive cycle[7]. Hence, we propose a computational technique to diagnose the battery characteristics directly through software without any environmental degradation[8].

Further, electrical characterization techniques have been analyzed on the data acquisition software i.e. LabVIEW by providing interfacing between hardware and software system[9]. Initially, a hardware prototype for testing a lithium-ion coin cell battery has been designed with the help of a simple electronic circuit. This hardware prototype was interlinked with data acquisition software through NI-6001 DAQ card[10]. DAQ card provides a basic function for voltage measurement and transferred hardware data to the real-time LabVIEW software for data analysis. The aim of these experiments is to optimize the current and voltage passed within the circuit during discharge condition and also for performing other safety tests.

Further, this paper is demonstrated as follows. Section 2 describes the model formulation of the lithium-ion battery and also explains the experimental setup in brief. In section 3, mechanical and electrical results are presented with the help of different software. Hence, the paper is summarized with the conclusion in section 5.

2. MODEL GEOMETRY AND FORMULATION

2.1 Lithium-ion battery model formulation

Computational analysis has been performed for this 1D model geometry. In this simulation study, we optimized the battery behavior for an electric vehicle in which we focused on three critical components i.e. specific drive cycle of the vehicle, temperature, and voltage of the battery[11]. These all three parameters have been monitored and act as an input to the control unit. This monitoring unit helps to demonstrate the state of charge (SOC) of the battery and also decides the battery charging (full) and discharging (empty) condition.

Firstly, the lithium-ion battery model has been formulated on the COMSOL software[12]. In this battery, lithium manganese oxide ($LiMn_2O_4$) is used as a positive electrode and graphite as a negative electrode. These both electrodes were separated with the help of separator (microporous polypropylene sheet Celgard 2400) and filled with liquid electrolyte ($LiPF_6$ EC:DMC-1:1 ratio) for the movement of ions during charging and discharging process. Furthermore, the outside layer of both positive and negative electrodes was connected with the current collector (CC) made up of aluminum and copper material.

In this formulation, firstly, we focused on the thermal characterization of the material, it has been analyzed by utilizing a heat transfer in the solid interface over a 1D geometry. In 1D geometry, a flat prismatic battery has been taken and allows heat transportation only in one direction. There were two conditions in the prismatic battery for characterized the thermal behavior i.e. thermal insulation and convective heat flux condition. Initially, we set the thermal insulation in the leftmost boundary of the battery which also represents the center of the battery. Similarly, heat flux condition is set in

the rightmost boundary which represents the heating surface of the battery. For air cooled the battery, the heat transfer coefficient is also applied roughly on the boundaries of the battery i.e. 10 W/ (m^2.K). There are some other parameters discussed in Table. I for thermal characterization techniques of the 1D battery model.

TABLE. I Thermal Properties For 1d Battery Model

Parameters	**$LiMn_2O_4$**	**Graphite**	**$LiPF_6$**	**Aluminum CC**	**Copper CC**
Thermal Conductivity(W(m.k))	3.4	1	0.15	400	238
Young Modulus (Pa)	$4.9*e^9$	$4.1*e^9$	-	$110*e^9$	$70*e^9$
Density (Kg/m^3)	3600	2260	940	8960	2700
Heat Capacity(J/Kg.k)	2500	881	1046	385	900

Additionally, a current load has analyzed for driving an electric vehicle. Hence, we imported a drive cycle data which contains C-rate vs time for this lithium-ion battery model. This drive cycle data may vary up to 20C, and 1C equal to 12A.

2.2 Lithium-ion battery model parameters

The model assembly has certain other parameters, which have been taken from research papers and parameters for designing a coin cell such as electrode thickness, the total thickness of the cell and many more are taken from a research laboratory. These are the operational parameters for simulation analysis. However, these can be changed according to the requirement of simulation analysis. These were the maximum number of parameters that identified theoretically in the lithium-ion battery model and further, other parameters were demonstrated in Table. II Moreover, the stoichiometry of both positive and negative electrode transfer coefficient and the electrolyte charge transfer coefficient have been taken as a known parameter from the literature. These physical properties are explained and optimized according to the simulation analysis. In the next section, the simulation results will be investigated and calculated.

TABLE. II Measured And Identified Parameters For The Lithium-Ion Battery Model Simulation

Parameter	**Units**	**Value**	**Description**
E_cell	V	3.9	Cell Potential
L_tot	m	$1.42*e^{-4}$	Cell total Thickness

epsl_sep	-	0.4	Porosity of separator
$T_{initial}$	K	328	Initial Temperature
T_{final}	K	395	Final Temperature
Maximum electrodeSOC	SOCmax	0.79	Operational SOC
Minimum electrode SOC	SOCmin	0.02	Operational SOC
Q_B	A*h	12	Battery Capacity
N_layers	-	25	No. of cell layers
kT_batt	L_batt/(L_pos/kT_pos+L_neg/kT_neg+L_pos_cc/kT_pos_cc+ L_neg_cc/kT_neg_cc+L_sep/kT_sep)	0.87451 W/(m·K)	Battery thermal conductivity
Rho_batt	(rho_pos*L_pos+rho_neg*L_neg+rho_pos_cc*L_pos_cc+rho_neg_cc*L_neg_cc+rho_sep*L_sep)/L_batt	2189.3 kg/m³	Battery density
Cp_batt	(Cp_pos*L_pos+Cp_neg*L_neg+Cp_pos_cc*L_pos_cc+Cp_neg_cc*L_neg_cc+Cp_sep*L_sep)/L_batt	1391.8 J/(kg·K)	Battery heat capacity
C	-	1	C-rate factor

3. OVERVIEW OF EXPERIMENT

For an electrical characterization of coin cell lithium-ion battery, we have to prepare a simple electronic circuit in which Li-ion battery bank has connected for optimizing the discharge curves with the help of commercial software i.e. LabVIEW.

3.1 Materials and equipment

For these experiments, we have six coin-cell batteries (max 3.5V) which are fabricated in a research laboratory, variable resistance (0-500 ohm), motor as a load, capacitors, three switches, wires, a protoboard, a digital multimeter, a personal computer connected with National Instrument 6001 DAQ card and commercial software LabVIEW with updated version[11].

3.2 Experiment setup

In this hardware, three electrical tests were performed for optimization of different types of discharge curves. figure. 1 represents a schematic circuit diagram of the hardware prototype in which, we demonstrate a circuit according to the test. Therefore, there were three working conditions in this hardware, firstly when switch SW1, SW4 is closed for monitoring the potential drop across the battery terminal as well as current passes across the resistor is measured through DAQ card channels with respect to time. In the second condition, Switch SW1, SW2 is closed for analyzing a short circuit current across the resistor through an analog input channel 2. Lastly, Switch SW1, SW3, SW4 are closed for monitoring the battery characteristics, when the capacitor bank is connected parallel across the circuit.

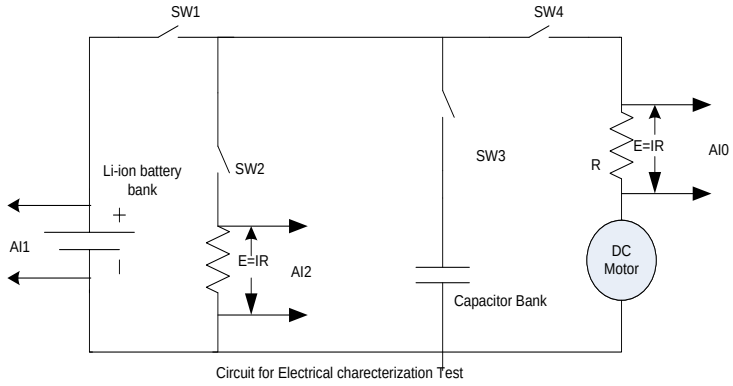


Fig. 1 Circuit for Hardware Prototype

When the voltage signal passes across the variable resistor terminals, then data has been recorded by two different input channels of the NI-6001 DAQ card [13]and further this device make connection to the PC for automatic data acquisition[14].

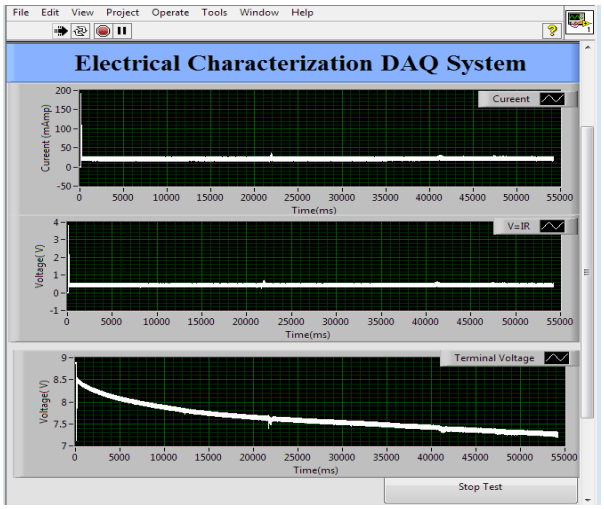


Fig. 2 Demonstrates the discharge curve on the LabVIEW Front Panel

We use LabVIEW software for data acquisition which helps to optimize the real-time physical data across various load conditions[14]. There are different in-built functions in this software such as logic components, graphics, a mathematical function, and many data acquisition icons which helps to perform automated measurements[15]. For writing a program we have two different panels on this software, one is a front panel which helps as an interface between the hardware and the PC[16]. It also helps for modifying the experimental condition and introduce virtual information across the panel i.e. it is called a virtual instrument (VI). The front panel helps to introduce the parameters of the hardware circuit such as voltage, current, resistor, experiment time and a number of channels used for recording the data. The second panel is used for designing a block diagram and also provide a proper platform for writing a program by using various icons and connectors. Figure. 2 demonstrates the discharge curve on a LabVIEW front panel.

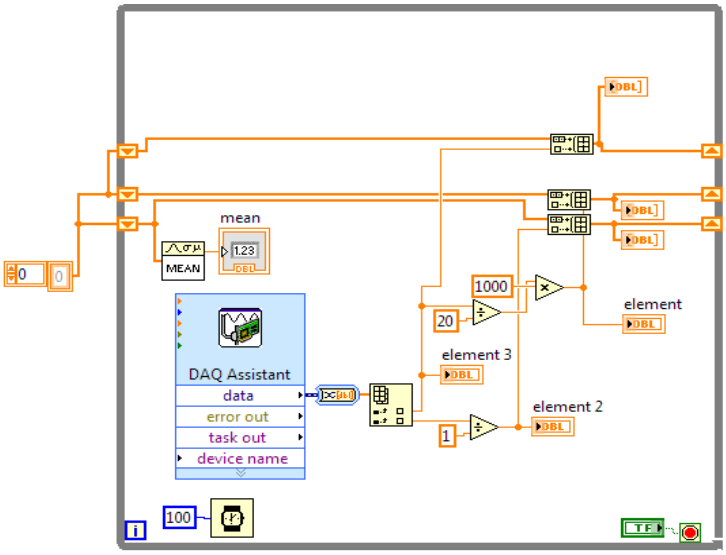


Fig. 3 Block Diagram of VI front panel

In figure. 3, we represent a block diagram for this VI front panel. This panel controls the sequence of execution and also use the LabVIEW sequence structure. Firstly, we

acquire the data points at the desired voltage channel and then we have to construct a time vector by using a loop structure at the desired rate. Lastly, before plotting a graph we have to separate the data array into two different vectors i.e. potential drop measured across the resistor and voltage signal. Figure. 4 represents a hardware prototype made up by 2032-coin cell Lithium-ion battery which was fabricated in a research laboratory.

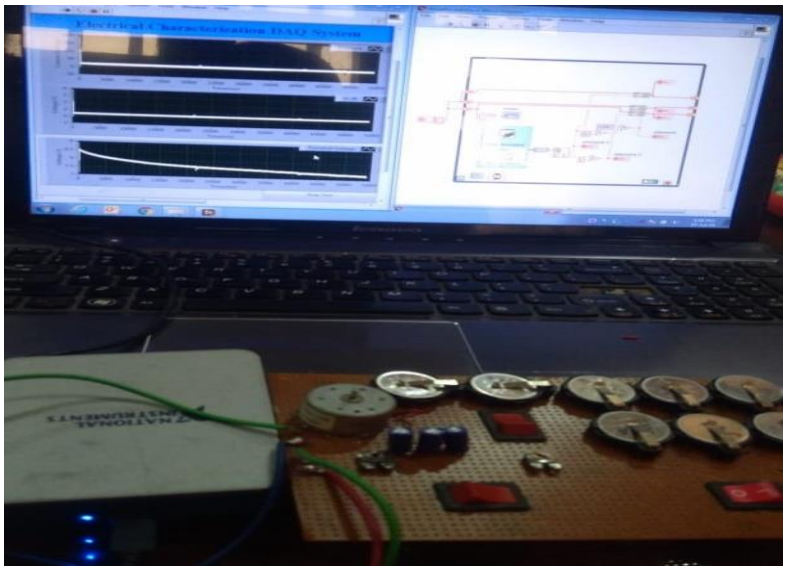


Fig. 4 Hardware Prototype for Electrical Characterization

Hence, both the theoretical model and hardware prototype setup have certain parameters which can change according to the requirements. Furthermore, in this paper, the experimental and simulation setup has been used to optimize the battery characteristics[17].

4. RESULT AND DISCUSSION

During battery characterization, the following tests have been performed to optimize the behavior of the Li-ion battery. All simulation results were analyzed at different boundary condition. Hence, all results were explained further with brief descriptions.

4.1 Mechanical Characterization

In this characterization, we have designed a coin cell battery in which a heat source generator used to raise the temperature (30^0C- 80^oC) of the battery through the separator. In this model, the convective heat flux boundary condition was set using homogenized thermal material properties across the battery. Initially, the temperature was at a constant rate then it starts increasing slowly. Figure. 5 represents the surface temperature at the positive electrode CC during the drive cycle. In this result, the surface temperature accumulated the heat inside the cell which could lead the thermal runway in a long time. If there would be a small variation in the temperature between surface and center then hot spots were reflected on the cell.

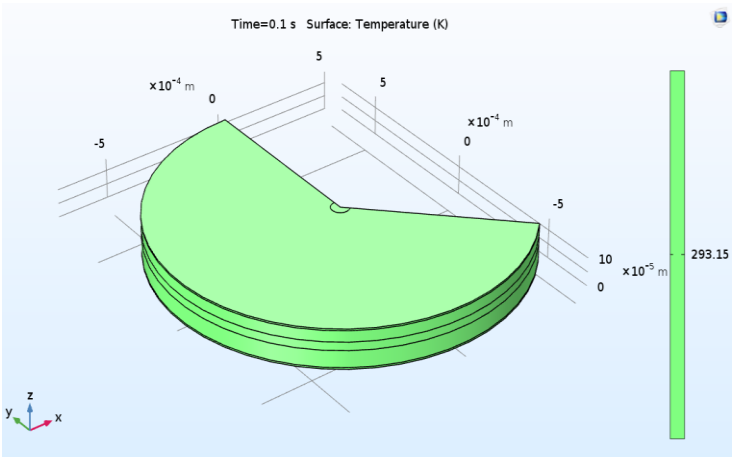


Fig. 5 Coin-Cell Surface Temperature

In figure. 6, plot represents the increase in temperature during the drive cycle. In this, the temperature of the battery on the surface as well as in the center of the coin cell had increased simultaneously otherwise it would lead to destroying the battery surface.

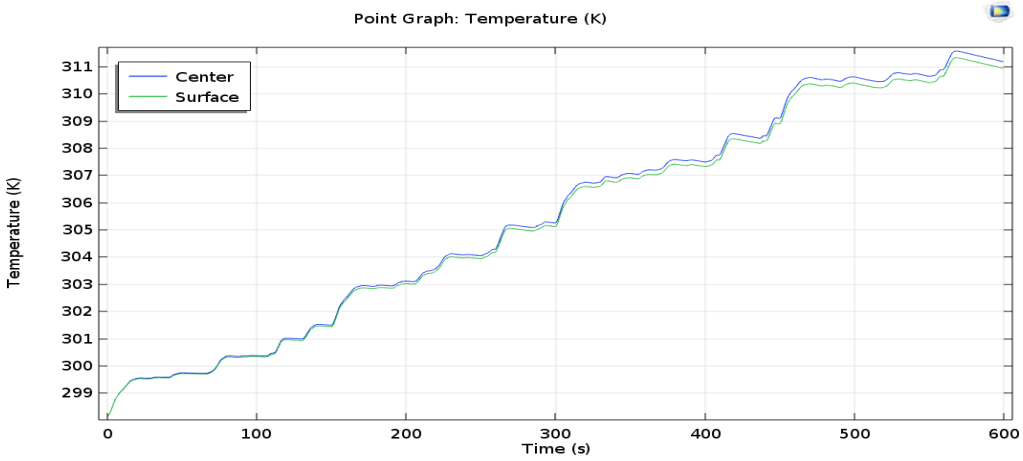


Fig. 6 Temperature of the Battery During Drive Cycle

During this thermal modeling, polarization characteristics of the coin cell battery have optimized because it gives an indication of internal resistance and it also explains the battery lifetime phenomenon. If there would be large polarization takes place across the battery than it reduces the battery life as well as cause heat generation. Hence, battery polarization should always be small. Figure. 7 represents the total polarization of the battery during the drive cycle.

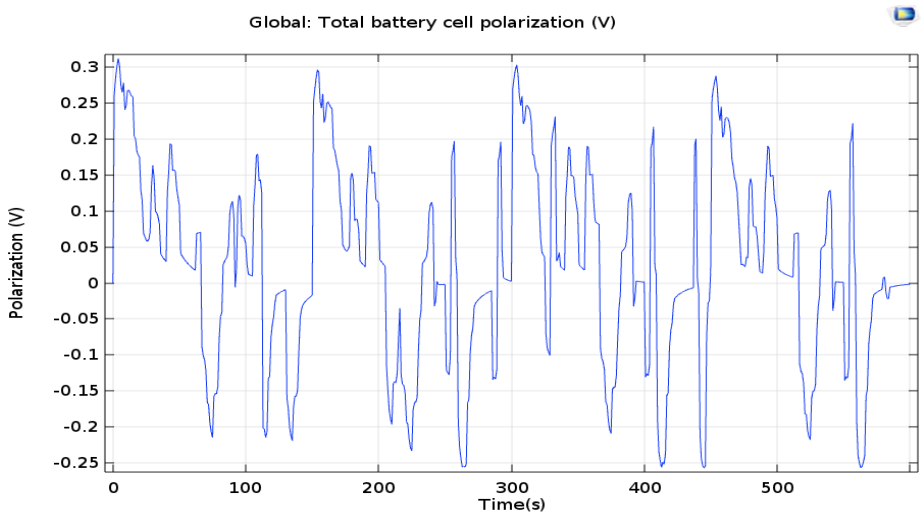


Fig. 7 Total Polarization During Drive Cycle

After this thermal characterization, we have to analyze the battery charging characterization. Initially, the cell voltage varies between 3.6V to 2.5V due to the open circuit voltage. But after a certain time when the drive cycle starts at certain C-rate then the potential of the positive electrode in the lithium-ion battery varies slightly more the negative electrode of the battery[18]. Figure.8 (a) represents a Cell voltage during the drive cycle.

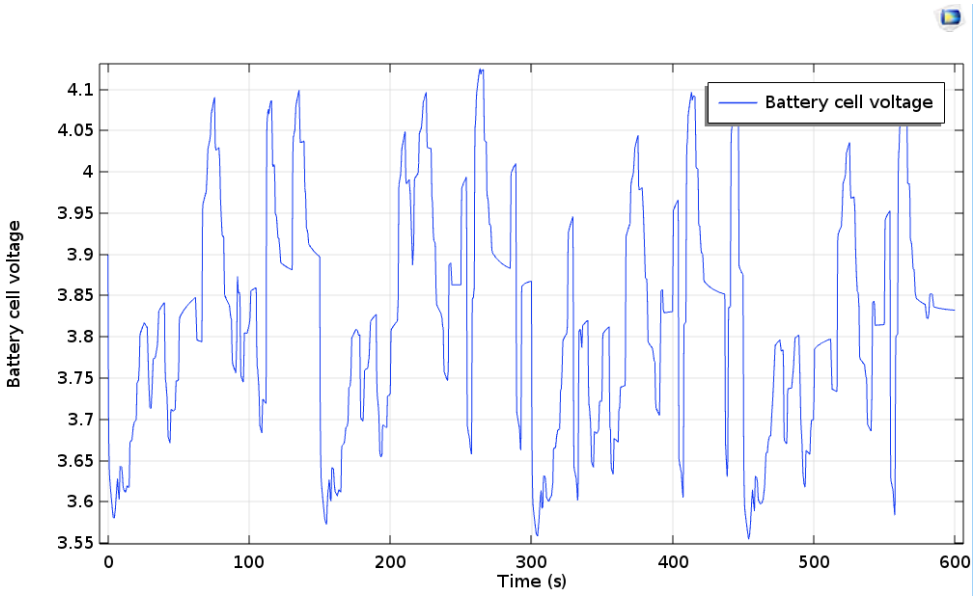


Fig. 8 (a) Cell Voltage during Drive Cycle

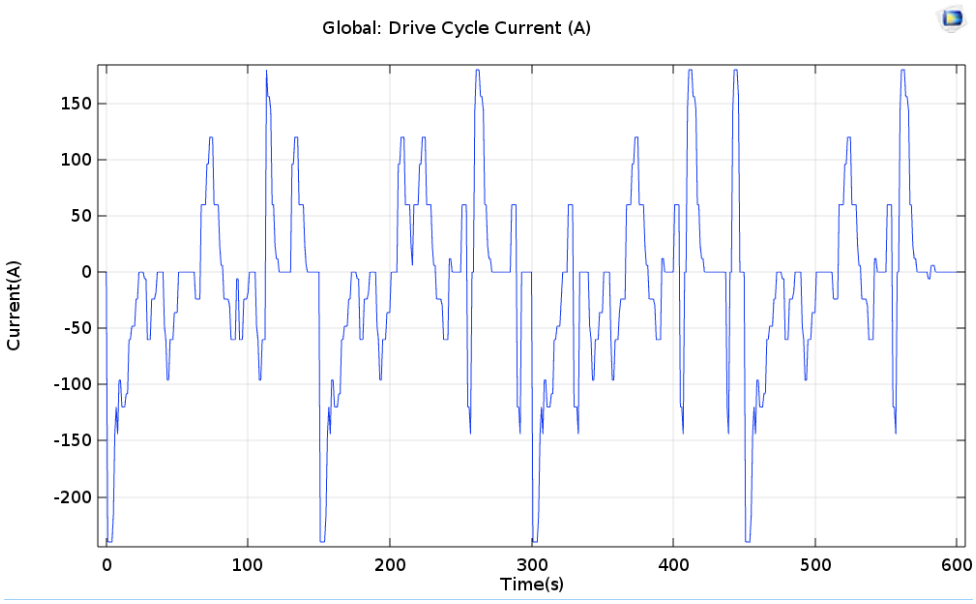


Fig. 8 (b) Drive Cycle Current on Load

Figure. 8 (b) represents the Drive Cycle Current on load. Battery voltage should always be in a limited range. Hence, we always operate a battery within a specified voltage span (4.2V to 3.3V) due to the battery safety concern.

4.2 Electrical Characterization

In this characterization, two types of test have been performed for monitoring the electrical performance of the battery bank i.e. short circuit test and battery bank connected parallel with a capacitor bank for improving the battery efficiency.

Short circuit test has been performed because it can be a major issue for real-time application because it provides high discharge current, a sudden drop in voltage and causes excess overheating. Figure.9 represents a short circuit test waveform, During this test, we monitored a sudden increase in current (500mA) across the resistor terminal of the battery. The resistor used in this test was very small i.e. 1.6 ohm and the potential drop across it was 0.9V. As there was a sudden sweep in voltage drop occurs across the battery due to a short circuit test.

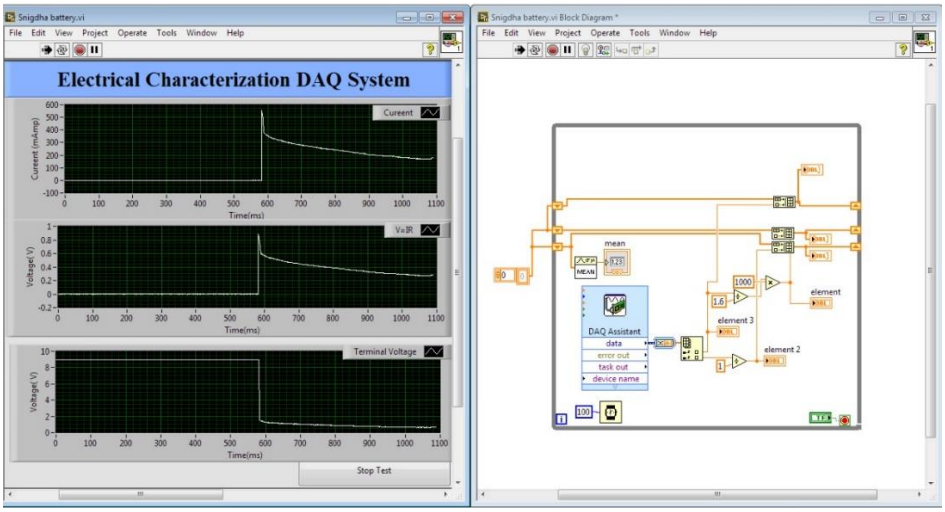


Fig. 9 Short Circuit Test Waveform

Now, we monitored a current and voltage waveform during normal condition. In this circuit, the load was connected to a battery bank to optimize the discharge characteristics of the battery bank. During this test, the current across the resistor terminal is increased up to 80mA and then stable at 20mA. Similarly, a potential drop across the 24Ω resistor is 0.5V and battery bank potential is 7.4V. Figure. 10 represents the current and voltage waveform without capacitor bank.

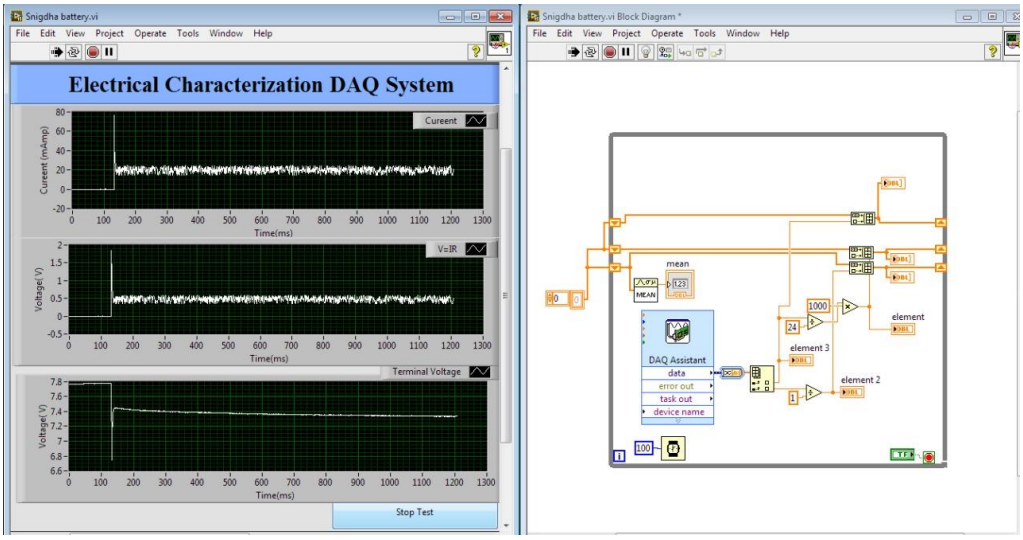


Fig. 10 Current And Voltage Waveform Without Capacitor Bank

Further, the three capacitors of 220 microfarads were connected parallel to form a capacitor bank and this capacitor bank was also connected parallel with a battery bank to improve the battery efficiency. During this test, we monitored a sudden rise in current across the resistor i.e. 148mA as we see in the figure. 11 current and voltage waveform with the capacitor bank.

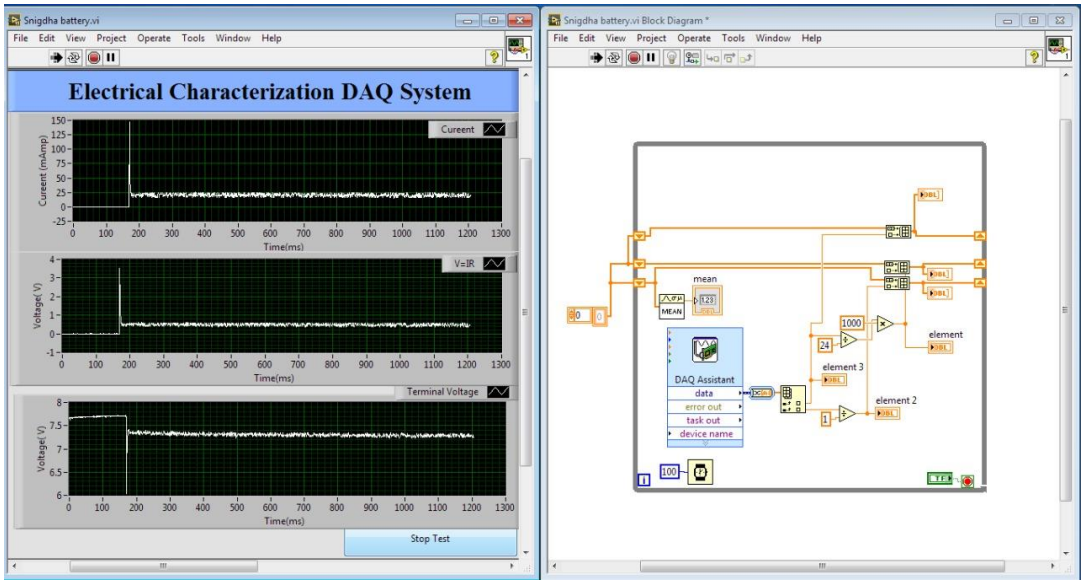


Fig. 11 Current And Voltage Waveform With Capacitor Bank

Hence, we optimized that as we connect a capacitor across the battery, then it provides high accelerating current across the circuit and then it becomes stable at 25mA. The voltage monitored across the battery bank was the same at 7.4V. So, It has been analyzed that the increase in current across the load would be delivered by a capacitor bank.

5. CONCLUSION

In this paper, theoretical and hardware simulation were successfully performed using Multiphysics COMSOL and LabVIEW software. We have performed two types of simulation for obtaining a mechanical and electrical characterization of a lithium-ion battery. In mechanical characterization, we optimized that Li-ion battery behavior during the drive cycle has good capacity and it has superior thermal behavior properties. Similarly, in electrical characterization test, we obtained a good electrical behavior of the Li-ion battery bank.

Acknowledgment

Authors are thankful to LIBT lab, Department of Applied physics from Delhi Technological University, New Delhi for providing support to carry out this simulation work.

References

1 Thierry Cassagneau and Janos H Fendler, Advanced Materials **10** (11), 877 (1998).

2 Xiaosong Hu, Fengchun Sun, and Yuan Zou, Energies **3** (9), 1586 (2010).

3 Rakesh Saroha, Amrish K Panwar, and Yogesh Sharma, Ceramics International **43** (7), 5734 (2017).

4 Linda Ager-Wick Ellingsen, Guillaume Majeau-Bettez, Bhawna Singh, Akhilesh Kumar Srivastava, Lars Ole Valøen, and Anders Hammer Strømman, Journal of Industrial Ecology **18** (1), 113 (2014); Xiaosong Hu, Fengchun Sun, Yuan Zou, and Huei Peng, presented at the Proceedings of the 2011 American Control Conference, 2011 (unpublished).

5 Rakesh Saroha, Amrish K Panwar, Yogesh Sharma, Pawan K Tyagi, and Sudipto Ghosh, Applied Surface Science **394**, 25 (2017); Victoria L McLaren, Anthony R West, Mitsuharu Tabuchi, Akiko Nakashima, Hikari Takahara, Hironori Kobayashi, Hikari Sakaebe, Hiroyuki Kageyama, Atsushi Hirano, and Yasuo Takeda, Journal of The Electrochemical Society **151** (5), A672 (2004).

6 Mingzhi Dai, Changhe Huo, Qiang Zhang, Karim Khan, Xingye Zhang, and Cai Shen, Advanced Theory and Simulations **1** (10), 1800023 (2018).

7 Farid Bahiraei, Maryam Ghalkhani, Amir Fartaj, and Gholam-Abbas Nazri, Applied Thermal Engineering **125**, 904 (2017).

8 Christian Hellwig, Seniz Sörgel, and Wolfgang G Bessler, Ecs Transactions **35** (32), 215 (2011).

9 Xiao-song Hu, Feng-chun Sun, and Yuan Zou, Journal of Central South University of Technology **18** (5), 1525 (2011).

10 Asan Gani and MJE Salami, presented at the Student conference on research and development, 2002 (unpublished).

11 Qiu Tang, Zhaosheng Teng, Siyu Guo, and Yaonan Wang, presented at the 2009 International Conference on Measuring Technology and Mechatronics Automation, 2009 (unpublished).

12 Long Cai and Ralph E White, Journal of Power Sources **196** (14), 5985 (2011).

13 Chih-Chiang Hua and Meng-Yu Lin, presented at the ISIE'2000. Proceedings of the 2000 IEEE International Symposium on Industrial Electronics (Cat. No. 00TH8543), 2000 (unpublished).

[14] Stephan Buller, Eckhard Karden, Andreas Lohner, and Rik W De Doncker, presented at the INTELEC-Twentieth International Telecommunications Energy Conference (Cat. No. 98CH36263), 1998 (unpublished).
[15] Anif Jamaluddin, Louis Sihombing, Agus Supriyanto, Agus Purwanto, and M Nizam, presented at the 2013 Joint International Conference on Rural Information & Communication Technology and Electric-Vehicle Technology (rICT & ICeV-T), 2013 (unpublished).
[16] Lorenzo Rambaldi, Enrico Bocci, and Fabio Orecchini, Applied Energy **88** (2), 442 (2011).
[17] Anif Jamaluddin, Anafi Nur'aini, Arif Jumari, and Agus Purwanto, International Journal of Power Electronics and Drive Systems **8** (2), 926 (2017).
[18] Dominic A Notter, Marcel Gauch, Rolf Widmer, Patrick Wager, Anna Stamp, Rainer Zah, and Hans-Jorg Althaus, (ACS Publications, 2010).

Research Article Vol. 2, No. 8 / 15 August 2019 / OSA Continuum 2353

Negative axicon tip micro-cavity with a polymer incorporated optical fiber temperature sensor

DNYANDEO PAWAR,[1,3] AJAY KUMAR,[1] RAJESH KANAWADE,[1] SAMIR MONDAL,[1] AND RAVINDRA K. SINHA[1,2,4]

[1] *CSIR-Central Scientific Instruments Organisation, Advanced Materials and Sensors Division (V4), Sector 30-C, Chandigarh – 160030, India*
[2] *TIFAC-CORE, Department of Applied Physics, Delhi Technological University, Bawana Road, Delhi-110042, India*
[3] *pawar.dnyandeo@gmail.com*
[4] *dr_rk_sinha@yahoo.com*

Abstract: A negative axicon micro-cavity polydimethylsiloxane (PDMS) filled Fabry-Perot interferometer (FPI) based sensor for accurate temperature sensing is proposed and demonstrated. The micro-cavity lengths of 130 µm, 160 µm, 240 µm and 260 µm are employed for temperature sensing from 27 °C to 80 °C. The sensing probe responses were measured in terms of wavelength shift resulting from the thermo-induced change in the FP cavity length and in the refractive index of the PDMS filled in the cavity. However, the effect of thermal expansion/contraction of PDMS is found to be more dominating over the change in its refractive index. The highest and least sensitivity of the order of 59 pm/ °C and 24 pm/ °C are observed for 130 µm and 260 µm cavity lengths. The linear relationship between the change in spatial frequency and cavity length with respect to temperature variation are also studied. These miniaturized and stable sensor probes are capable of measuring small change in temperature variation with high accuracy and sensitivity and can be used for remote sensing measurements.

© 2019 Optical Society of America under the terms of the OSA Open Access Publishing Agreement

1. Introduction

Accurate temperature sensing is being keen interest for researcher's due to its various applications in bioscience, pharmaceutical and food science etc. To fulfill their requirements a highly sensitive, stable, biocompatible, reliable and durable sensor is always on anvil. Fiber optic-based sensors could meet such demands and proved to be suitable for such measurement due to its inert nature, immune to electromagnetic interference (EMI), and distributed sensing ability. Various fiber optic configurations were proposed and fabricated for temperature measurement including Mach-Zehnder [1], Sagnac loop [2], and Fabry-Perot interferometric (FPI) [3]. Certain geometries including thin film inside the single mode fiber (SMF) [4], polymer coating at the distal end of the fiber [5,6], long period gratings (LPGs) [7], photonic crystal fiber-Fiber Bragg gratings (PCF-FBGs) [8], femto-second laser micromachining [9], hollow core PCF-SMF [10], and liquid filled PCF [11] were demonstrated for the temperature sensing.

Among the various interferometers, FPI based geometry could be the most suitable for temperature measurement due to its simplicity, high sensitivity and remote sensing ability. Some advancement has been done in the field of FPI-based tip sensors such as cantilever-based [12], polymer coated [13,14], monolithic photonic crystal based [15] and nano-patterned based [16] for temperature sensing. However, some of these sensors possesses low thermal expansion coefficient, low mechanical stability and durability and thus limits it uses in on-field temperature sensing applications. Therefore, to overcome these limitations there is a need of highly sensitive, mechanically stable and a reliable sensor which could fulfill the requirement of the industry.

Considering the above limitations in fabricating the FPI cavity at the tip of an optical fiber, we have proposed and fabricated a negative axicon micro-cavity filled with Polydimethylsiloxane

#367545 https://doi.org/10.1364/OSAC.2.002353
Journal © 2019 Received 13 May 2019; revised 27 Jun 2019; accepted 5 Jul 2019; published 23 Jul 2019

OSA CONTINUUM

(PDMS) for temperature measurement. PDMS is chosen for sensing due to its characteristics such as biocompatible, non-toxic, optically transparent and possess very high thermal expansion coefficient of the order of 4.66×10^{-4}/ °C [17]. Hence, PDMS filled negative axicon cavity could be a promising sensing platform for temperature measurement. This type of cavity is existing inside the optical fiber and thus mechanically stable. As the PDMS is filled inside the cavity, there will be relatively less impact of humidity on it and becomes mainly sensitive to the temperature variation. This type of cavity possesses large cavity volume, so that we can tune the cavity length accordingly. Therefore, this sensor possesses many interesting features like cost-effectiveness, simple to construct and mechanically stable and thus may improve the accuracy of the temperature measurement.

2. Fabrication and experiment

The schematic of a negative axicon tip micro-cavity fabrication of the PDMS (SYLGARD 184) filling inside an air-cavity is shown in Fig. 1. The sensor tips were fabricated by using simple hydrofluoric (HF) acid treatment in which the Cladding Mode Offset Photosensitive Single Mode Fiber (CMOP-SMF) (highly Ge doped optical fiber with core diameter ~3.5 µm, Nufern, USA) was immersed in HF acid solution (concentration 48%) as shown in Fig. 1(a). The total etching time of fabricating the air-cavity inside the fiber was around 40 minutes. The core of the highly Ge doped fiber etched faster than the cladding and hence it forms an air-cavity inside the fiber. The quality of the fabricated micro-cavity in terms of surface roughness and uniform etching was confirmed through beam profile before PDMS filling [18]. The cavity length can be controlled by varying the etching time. As shown in Fig. 1(b), the PDMS was taken at the tip of a positive axicon fiber and filled inside an air cavity of a negative axicon fiber with the help of high precision XYZ translation stage (561D-XYZ ULTRAlign 3 axis stage, Newport). It observed that the lengths of the PDMS filled micro-cavities were of the order of ~130 µm, ~160 µm, ~240 µm and ~260 µm respectively.

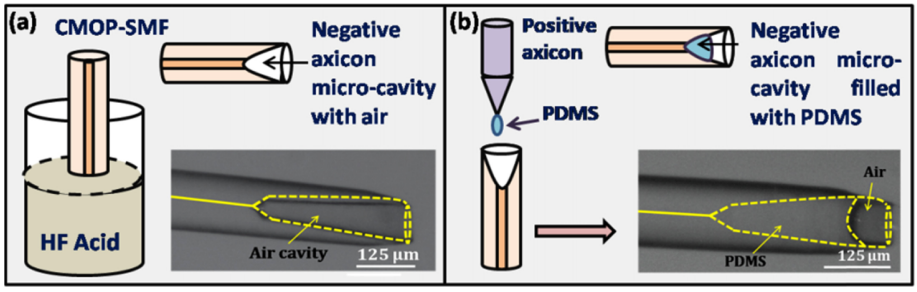


Fig. 1. Schematic of negative axicon micro-cavity fabrication (a) etching (inset shows the fabricated probe) and (b) PDMS filing (inset shows the micro-graph of fabricated sensor probe).

The schematic of the proposed temperature sensor set up is shown in Fig. 2. It consists of an Amplified Spontaneous Emission (ASE) light source (1525-1565 nm, Output power 28 mW, Optolink Corp. Ltd., Hong Kong), a circulator (SMF-28, 60:40 ratio, Optolink Corp. Ltd., Hong Kong), an Optical Spectrum Analyzer (OSA) (AQ6370C, wavelength resolution 20 pm, 600-1700 nm, Yokogawa, Japan) and a temperature oven. The negative axicon filled with PDMS sensor was spliced at the distal end of the circulator. The light from PDMS-Air interface reflects and gets coupled in to the OSA to generate an interference pattern. The temperature response of all sensor probes were measured by placing it inside the temperature oven with gradually increasing the temperature from 27 °C to 80 °C with step of 10 °C. During each measurement,

OSA CONTINUUM

the temperature of the oven was verified by placing a thermometer inside the oven. With respect to temperature variation, the corresponding interference spectrums were measured and studied for further analysis.

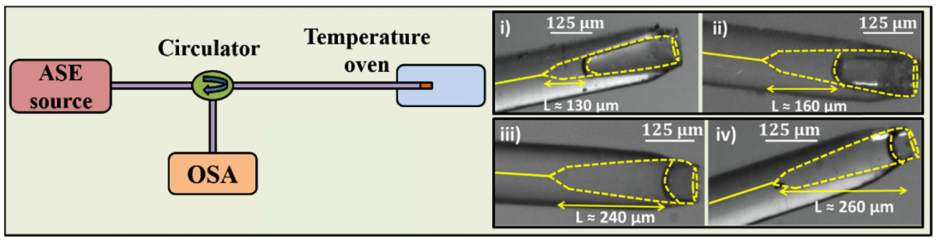


Fig. 2. Schematic diagram of an experimental set up of FPI-based PDMS filled micro-cavity for temperature sensing. Inset shows micrographs of micro-cavity lengths of i) 130 µm, ii) 160 µm, iii) 240 µm and iv) 260 µm respectively.

3. Theoretical discussion

Due to temperature effect, the properties of the PDMS changes which results into the change in cavity length. The optical path length (OPL) inside the FP cavity for round trip and free spectral range (FSR) is given by [4,19]:

$$l_{OPL} = 2nL \tag{1}$$

where n is the refractive index of the FP cavity and L is the length of the FP cavity. The change in OPL due to temperature variation inside the FP cavity is the function of the refractive index of PDMS and the cavity length is given by [4,10]:

$$\Delta l_{OPL} = 2(n\frac{dL}{dT} + L\frac{dn}{dT})\Delta T = l_{OPL}(\alpha + \beta)\Delta T \tag{2}$$

where α is the thermal expansion coefficient, β is the thermo-optics coefficient and ΔT is the change in temperature.

Furthermore, the change in length of the cavity either due to thermal expansion or contraction of PDMS, which induces change in dip wavelength ($\Delta\lambda_{dip}$) of an interference spectrum is given by [4,13]:

$$\Delta\lambda_{dip} = \frac{2n\Delta L}{m} = \frac{\Delta L\lambda_{dip}}{L} \tag{3}$$

Similarly, change in the refractive index of PDMS further induces the dip wavelength shift in an interference spectrum. Thus, the change in wavelength shift in terms of temperature variation is given by [10]:

$$\Delta\lambda = \frac{\lambda\frac{dn}{dT}}{n}\Delta T = \gamma_T\Delta T \tag{4}$$

The temperature also induces change in spatial frequency of wavelength spectrum and therefore, the spatial frequency shift due to temperature variation can be calculated by [19]:

$$\Delta\xi = \frac{1}{\lambda^2}\Delta l_{OPL} \tag{5}$$

where λ is the laser source wavelength and $\Delta\xi$ is the amount of spatial frequency shift due to temperature variation.

OSA CONTINUUM

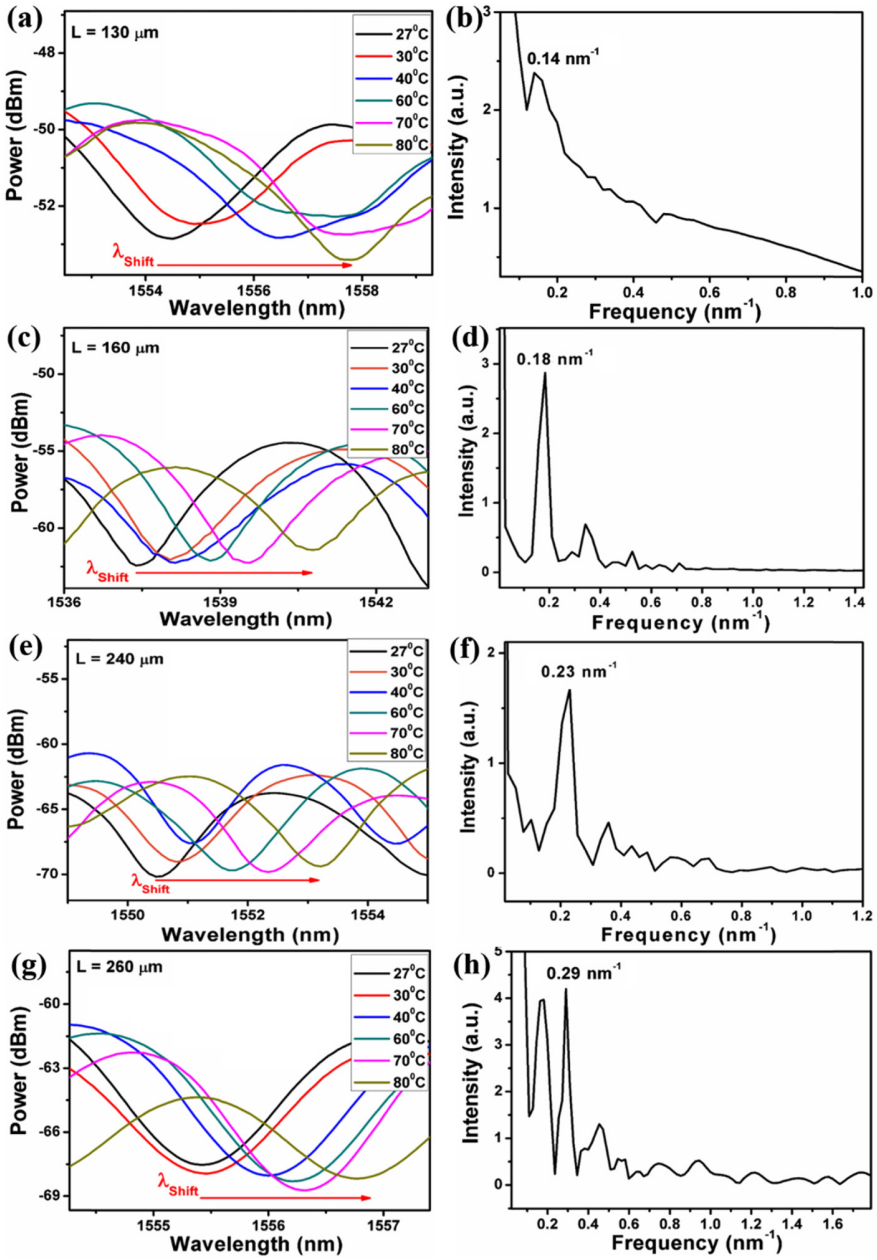


Fig. 3. The spectral response and its spatial frequency spectrum obtained by taking the FFT for each PDMS filled micro-cavity FPI with temperature variation from 27 °C to 80 °C for each cavity length of (a, b) for 130 μm, (c, d) for 160 μm, (e, f) for 240 μm and (g, h) for 260 μm respectively.

4. Results and discussion

The reflection spectrums of all sensor probes were recorded on the OSA in the temperature range from 27 °C to 80 °C. Figure 3 (a, c, e, g) shows that as the temperature increases from 27 °C

OSA CONTINUUM

to 80 °C the corresponding dip wavelength shift also increases. The dip wavelength shift was attributed due to the change in the properties of the PDMS in terms of change in cavity length resulting from the temperature effect. To understand the behavior of an interference fringes in detail, the spatial frequencies were calculated by applying the Fast Fourier Transform (FFT) for all micro-cavity of lengths 130 µm, 160 µm, 240 µm and 260 µm as shown in Fig. 3 (b, d, f, h). The theoretical FSR [13] value was calculated by using equation as $FSR = \frac{\lambda^2}{l_{OPL}}$ and found around 6.79 nm, 5.46 nm, 4.14 nm and 3.45 nm which corresponds to the theoretical spatial frequencies around 0.14 nm^{-1}, 0.18 nm^{-1} 0.24 nm^{-1} and 0.28 nm^{-1} for cavity length of 130 µm, 160 µm, 240 µm and 260 µm respectively. Experimentally, the spatial frequency peaks in FFT spectrums were observed around 0.14 nm^{-1}, 0.18 nm^{-1}, 0.23 nm^{-1} and 0.29 nm^{-1} as shown in Fig. 3 (b, d, f, h), which are nearly match with the theoretical values. Thus, the theoretical and experimental study confirms that the FP cavity is present inside the fiber.

The change in dip wavelength shift with respect to change in temperature variation is shown in Fig. 4, which shows the linear relation of dip wavelength shift with temperature variation from 27 °C to 80 °C for micro-cavity of lengths 130 µm, 160 µm, 240 µm and 260 µm respectively. The observed sensitivity of the fabricated sensors in the heating and cooling mode from temperature range of 27 °C to 80 °C were observed around 59.44 pm/ °C and 58.56 pm/ °C, 50.50 pm/ °C and 45.85 pm/ °C, 44.01 pm/ °C and 36.09 pm/ °C, 24.09 pm/ °C and 22.97 pm/ °C for micro-cavity lengths of 130 µm, 160 µm, 240 µm and 260 µm respectively. The approximate maximum error observed in the measurement were of the order of ± 0.54 nm, ±0.16 nm, ±0.16 nm and ± 0.16 nm for micro-cavity lengths of 130 µm, 160 µm, 240 µm and 260 µm respectively.

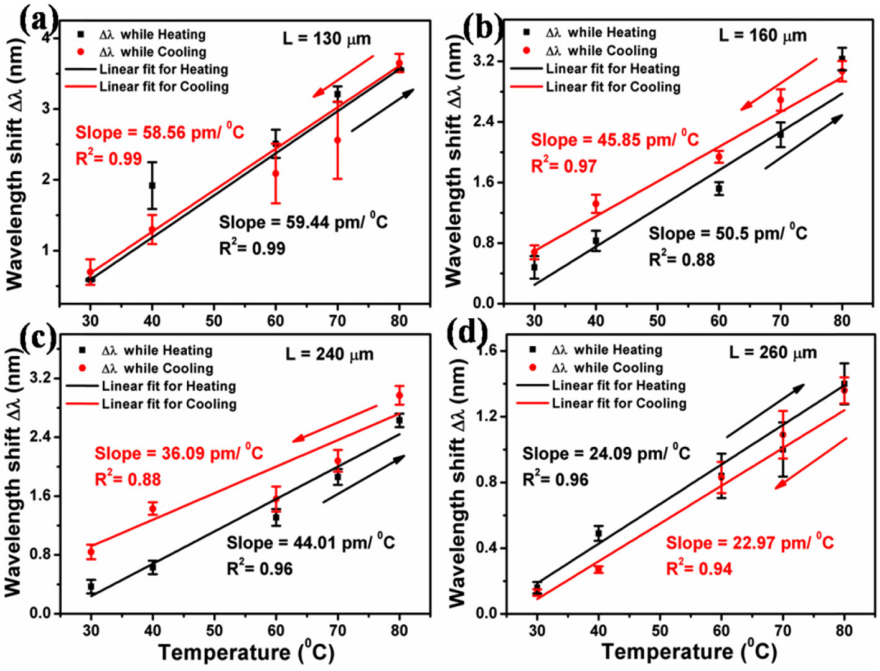


Fig. 4. Relationship between dip wavelength shift Vs temperature in heating and cooling mode for micro-cavity lengths of a) 130 µm, b) 160 µm, c) 240 µm and d) 260 µm respectively within temperature range of 27 °C to 80 °C.

In cooling mode, all sensors shown linear response with small systematic deviation compared with the results obtained from the heating mode. The reason for systematic deviation during

OSA CONTINUUM

cooling mode can be attributed due to the change in temperature gradient formed inside the cavity. The structure of the micro-cavity allows heating and cooling of the cavity in one direction only. It means during heating mode, the temperature gradient flows from air to PDMS and PDMS to air during cooling mode. The unequal temperature gradient formed inside the cavity during heating and cooling mode could be the reason for this systematic error in the measurement. Such systematic error was observed in three repeated experiment followed the similar trend during heating and cooling mode. However, overall sensitivity for all sensors in heating and cooling mode is found nearly of the same order.

The sensing response of the fabricated probes were tested in the temperature range from 27 °C to 80 °C. The limit of detection (LOD) of these sensors were observed around 0.33 °C, 0.39 °C, 0.45°C and 0.83 °C for cavity lengths of 130 µm, 160 µm, 240 µm and 260 µm respectively and it was limited due to the OSA resolution (wavelength resolution ≈ 20 pm). The response and recovery time of the sensors were around 8-10 seconds. From Eq. (3), it is confirmed that the wavelength shift change was large in case of 130 µm followed by 160 µm, 240 µm and 260 µm cavity lengths respectively. Theoretically, the PDMS thermally expands or contract due to thermal expansion effect and thermo-optic effect and hence it induces the wavelength shift in an interference spectrum. Such expansion/contraction can be confirmed from the cavity length changes which results in change in the sensitivity [17].

As discussed, the PDMS showed good characteristics relation with external temperature variation measured from the thermal and optical effect corresponding to the change in the cavity length. To understand the effect of temperature on the refractive index of PDMS at different temperature range, we measured the actual refractive index of PDMS in temperature variation from 27 °C to 80 °C by using digital refractometer (ATAGO, Rx-7000i, Japan) (shown in Fig. 5(a)). The linear response was observed with change in refractive index value of PDMS from 1.4104 to 1.3915 RIU in the applied temperature range from 27 °C to 80 °C. Thus, as discussed in theory the change in PDMS refractive index with temperature variation is also contributing into the change in wavelength shift.

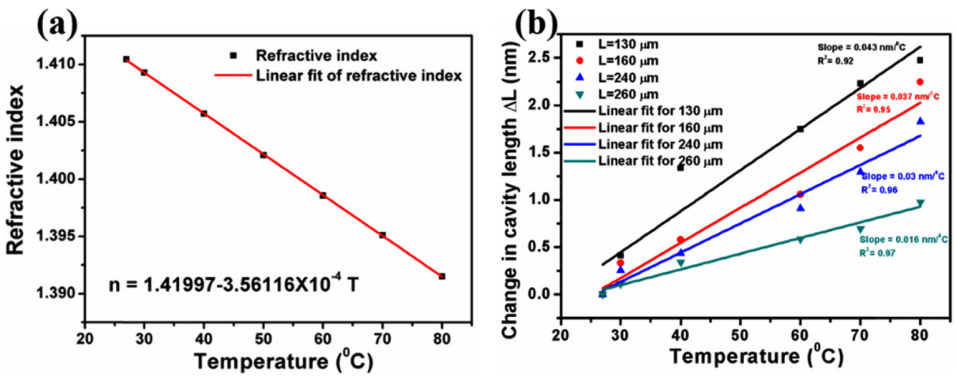


Fig. 5. (a) Refractive index variation of PDMS with change in temperature, (b) Change in length of micro-cavity for 130 µm, 160 µm, 240 µm and 260 µm in temperature variation from 27 °C to 80 °C.

The change in cavity length in terms of either due to thermal expansion or contraction was calculated from Eq. (3). Figure 5(b) shows the change in PDMS filled micro-cavity lengths due to temperature variation from 27 °C to 80 °C (during heating mode) for cavity lengths of 130 µm, 160 µm, 240 µm and 260 µm respectively. All sensing probes show the linear response with temperature variation. The corresponding sensitivity values were observed around 0.043 nm/ °C,

OSA CONTINUUM

0.037 nm/ °C, 0.030 nm/ °C and 0.016 nm/ °C in temperature range from 27 °C to 80 °C for the sensing probes of cavity lengths of 130 µm, 160 µm, 240 µm and 260 µm respectively.

Figure 6 shows the spatial frequency shift with temperature variation for different cavity lengths. The maximum spatial frequency shift of around 0.0020/nm, 0.0025/nm, 0.0037/nm and 0.0041/nm were observed with total temperature change 53 °C (temperature range from 27 °C to 80 °C) for cavity lengths of 130 µm, 160 µm, 240 µm and 260 µm respectively. Therefore, corresponding variation in round-trip OPL induced due to the temperature variation were calculated around 4.80 µm, 6.01 µm, 8.88 µm and 9.85 µm for the cavity lengths of 130 µm, 160 µm, 240 µm and 260 µm respectively.

Though, the effect of temperature on the PDMS through change in OPL study needs to be understood. First, we assume that the change in wavelength shift is induced due to the thermal expansion of PDMS and calculated from Eq. (2) modified as, $\mathbf{\Delta} l_{OPL} = l_{OPL}\alpha\mathbf{\Delta}T$ and obtained as 2.5×10^{-4}/ °C. Therefore, the theoretical value of thermal expansion is 4.66×10^{-4}/°C, which is close to the experimentally obtained value [17,20]. Secondly, consider the temperature variation rather than thermal expansion; we obtained thermo-optic coefficient from Eq. (2) as, $\mathrm{L}\mathbf{\Delta}\mathrm{n} = l_{OPL}\beta\mathbf{\Delta}T$ and obtained thermo-optic coefficient as 7.1×10^{-1}/ °C. However, actual thermo-optic coefficient of PDMS is reported as -4.5×10^{-4}/ °C [21], which is much smaller than the obtained value. Therefore, the thermal expansion of PDMS is more dominating over the change in refractive index due to temperature variation.

The obtained sensitivity is much higher than the sensors reported on FBGs [31] and PCF [8] with comparatively lower fabrication cost. The comparison of performance between the proposed sensor with various reported optical fiber-based sensors are tabulated in Table 1. There is scope

Table 1. The sensing performance of various developed fiber optic sensors with temperature.

Sensing material	Configuration/Mode	Sensitivity	Measurement range	Ref.
NOA65 Polymer	FP cavity / Reflection mode	249 pm/ °C	40 to 90 °C	[4]
PDMS	FP cavity / reflection mode	0.13 dB/ °C	22 to 60 °C	[17]
Teflon	microfiber coil resonator	6 pm/ °C	23 to 46 °C.	[22]
PVA	FP cavity / Reflection mode	−6.14 pm/ °C	10 to 50 °C	[23]
Gold coated membrane of Parylene C	FP cavity / Reflection mode	6.1 nm/ °C	26 to 50 °C	[24]
Toluene filled SC-PCF	Transmission mode	−6.02 nm/ °C	−80 to 90 °C	[25]
CdSe-ZnS core-shell QD doped PMMA microfiber	Transmission mode	58.5 pm/ °C	25 to 48 °C	[26]
Four core fiber -FBG	Transmission mode	194 pm/ °C	0 to 60 °C	[27]
Silicon cavity	FP cavity / Reflection mode	84.6 pm/ °C	20 to 100 °C	[28]
Graphene coated microfiber	Transmission mode	0.1052 dB/ °C	30 to 80 °C	[29]
SMF-FBG-MMF	Mach-Zehnder interferometer / Transmission mode	52 pm/ °C	25 to 69 °C	[30]
PDMS filled negative axicon tip	FP cavity / Reflection mode	59 pm/ °C	27 to 80 °C	This work

OSA CONTINUUM

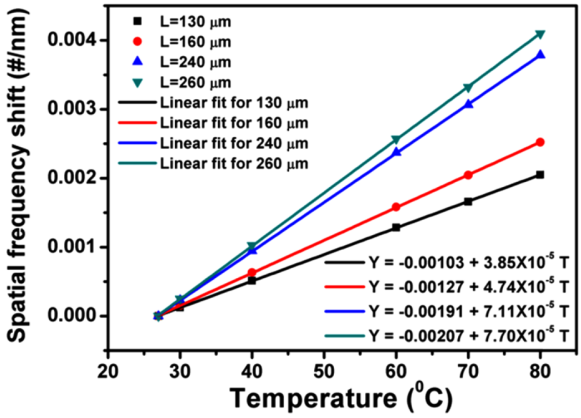


Fig. 6. The spatial frequency shifts Vs change in temperature from 27 °C to 80 °C for cavity lengths of 130 µm, 160 µm, 240 µm and 260 µm respectively.

to enhance the sensitivity further by changing the length of PDMS cavity. The systematic error can be removed by further changing the thickness of the film to make it more sensitive with optimized parameters.

5. Conclusion

The PDMS filled negative axicon tip microcavity-based sensors were proposed and demonstrated for temperature sensing. Due to temperature effect, PDMS thermally expands which changes the cavity length and further induced the wavelength shift in an interference spectrum. The sensors exhibited the sensitivity of the order of 59.44 pm/ °C for 130 µm; 50.50 pm/ °C for 160 µm; 44.01 pm/ °C for 240 µm and 24.09 pm/ °C for 260 µm cavity lengths in the temperature range of 27 °C to 80 °C. We believe that the favorable characteristics of this sensor such as simple, compact, sensitive, and biocompatible nature makes it ideal for temperature sensing for various industrial applications.

Funding

Council of Scientific and Industrial Research (CSIR) (HRDG/CSIR-Nehru PDF/EN, ES&PS/EMR-1/04/2018); Science and Engineering Research Board (SERB) (EMR/2015/001909).

References

1. P. Lu, L. Men, K. Sooley, and Q. Chen, "Tapered fiber Mach–Zehnder interferometer for simultaneous measurement of refractive index and temperature," Appl. Phys. Lett. **94**(13), 131110 (2009).
2. P. Zu, P. L. So, and C. C. Chan, "An Ultrahigh Sensitivity Point Temperature Sensor Based on Fiber Loop Mirror," IEEE J. Sel. Top. Quantum Electron. **23**(2), 274–277 (2017).
3. J. Mathew, O. Schneller, D. Polyzos, D. Havermann, R. M. Carter, W. N. MacPherson, D. P. Hand, and R. R. J. Maier, "In-Fiber Fabry–Perot Cavity Sensor for High-Temperature Applications," J. Lightwave Technol. **33**(12), 2419–2425 (2015).
4. B. Sun, Y. Wang, J. Qu, C. Liao, G. Yin, J. He, J. Zhou, J. Tang, S. Liu, Z. Li, and Y. Liu, "Simultaneous measurement of pressure and temperature by employing Fabry-Perot interferometer based on pendant polymer droplet," Opt. Express **23**(3), 1906–1911 (2015).
5. X. L. Tan, Y. F. Geng, X. J. Li, Y. L. Deng, Z. Yin, and R. Gao, "UV-Curable Polymer Microhemisphere-Based Fiber-Optic Fabry–Perot Interferometer for Simultaneous Measurement of Refractive Index and Temperature," IEEE Photonics J. **6**(4), 1–8 (2014).
6. O. Arrizabalaga, G. Durana, J. Zubia, and J. Villatoro, "Accurate microthermometer based on off center polymer caps onto optical fiber tips," Sens. Actuators, B **272**, 612–617 (2018).
7. L. Alwis, T. Sun, and K. T. V. Grattan, "Optical Fibre Refractive Index Sensor in a Hybrid Fibre Grating Configuration," Procedia Eng. **120**, 11–14 (2015).

OSA CONTINUUM

8. B. Dong, J. Hao, C.-y. Liaw, B. Lin, and S. C. Tjin, "Simultaneous strain and temperature measurement using a compact photonic crystal fiber inter-modal interferometer and a fiber Bragg grating," Appl. Opt. **47**(10), 1668–1672 (2008).
9. Z. L. Ran, Y. J. Rao, W. J. Liu, X. Liao, and K. S. Chiang, "Laser-micromachined Fabry-Perot optical fiber tip sensor for high-resolution temperature-independent measurement of refractive index," Opt. Express **16**(3), 2252–2263 (2008).
10. H. Y. Choi, K. S. Park, S. J. Park, U.-C. Paek, B. H. Lee, and E. S. Choi, "Miniature fiber-optic high temperature sensor based on a hybrid structured Fabry–Perot interferometer," Opt. Lett. **33**(21), 2455–2457 (2008).
11. A. Villalba and J. C. Martín, "Interferometric temperature sensor based on a water-filled suspended-core fiber," Opt. Fiber Technol. **33**, 36–38 (2017).
12. J. Li, F. Albri, J. N. Sun, M. M. Miliar, R. R. J. Maier, D. P. Hand, and W. N. MacPherson, "Fabricating optical fibre-top cantilevers for temperature sensing," Meas. Sci. Technol. **25**(3), 035206 (2014).
13. C. Wang, G. Yan, Z. Lian, X. Chen, S. Wu, and S. He, "Hybrid-cavity fabry-perot interferometer for multi-point relative humidity and temperature sensing," Sens. Actuators, B **255**, 1937–1944 (2018).
14. U. Sampath, D. Kim, H. Kim, and M. Song, "Cryogenic Temperature Sensor Based on Fresnel Reflection From a Polymer-Coated Facet of Optical Fiber," IEEE Sens. J. **18**(9), 3640–3644 (2018).
15. I. Jung, B. Park, J. Provine, R. T. Howe, and O. Solgaard, "Highly Sensitive Monolithic Silicon Photonic Crystal Fiber Tip Sensor for Simultaneous Measurement of Refractive Index and Temperature," J. Lightwave Technol. **29**(9), 1367–1374 (2011).
16. M. Consales, A. Ricciardi, A. Crescitelli, E. Esposito, A. Cutolo, and A. Cusano, "Lab-on-Fiber Technology: Toward Multifunctional Optical Nanoprobes," ACS Nano **6**(4), 3163–3170 (2012).
17. I. Hernández-Romano, M. A. Cruz-Garcia, C. Moreno-Hernández, D. Monzón-Hernández, E. O. López-Figueroa, O. E. Paredes-Gallardo, M. Torres-Cisneros, and J. Villatoro, "Optical fiber temperature sensor based on a microcavity with polymer overlay," Opt. Express **24**(5), 5654–5661 (2016).
18. K. Vairagi, R. A. Minz, S. Kaur, D. Kumbhakar, S. Paul, U. Tiwari, R. K. Sinha, J. Fick, and S. K. Mondal, "Deep Seated Negative Axicon in Selective Optical Fiber Tip and Collimated Bessel Beam," IEEE Photonics Technol. Lett. **29**(10), 786–789 (2017).
19. H. Y. Choi, G. Mudhana, K. S. Park, U.-C. Paek, and B. H. Lee, "Cross-talk free and ultra-compact fiber optic sensor for simultaneous measurement of temperature and refractive index," Opt. Express **18**(1), 141–149 (2010).
20. C. He, J. Fang, Y. Zhang, Y. Yang, J. Yu, J. Zhang, H. Guan, W. Qiu, P. Wu, J. Dong, H. Lu, J. Tang, W. Zhu, N. Arsad, Y. Xiao, and Z. Chen, "High performance all-fiber temperature sensor based on coreless side-polished fiber wrapped with polydimethylsiloxane," Opt. Express **26**(8), 9686–9699 (2018).
21. C. Markos, K. Vlachos, and G. Kakarantzas, "Bending loss and thermo-optic effect of a hybrid PDMS/silica photonic crystal fiber," Opt. Express **18**(23), 24344–24351 (2010).
22. Y. Chen, F. Xu, and Y.-Q. Lu, "Teflon-coated microfiber resonator with weak temperature dependence," Opt. Express **19**(23), 22923–22928 (2011).
23. S. Wu, G. Yan, Z. Lian, X. Chen, B. Zhou, and S. He, "An open-cavity Fabry-Perot interferometer with PVA coating for simultaneous measurement of relative humidity and temperature," Sens. Actuators, B **225**, 50–56 (2016).
24. M. Llera, T. Aellen, J. Hervas, Y. Salvadé, P. Senn, S. Le Floch, and H. Keppner, "Liquid-air based Fabry-Perot cavity on fiber tip sensor," Opt. Express **24**(8), 8054–8065 (2016).
25. J. Ma, H. H. Yu, X. Jiang, and D. S. Jiang, "High-performance temperature sensing using a selectively filled solid-core photonic crystal fiber with a central air-bore," Opt. Express **25**(8), 9406–9415 (2017).
26. N. Irawati, S. W. Harun, H. A. Rahman, S. S. Chong, N. A. Hamizi, and H. Ahmad, "Temperature sensing using CdSe quantum dot doped poly(methyl methacrylate) microfiber," Appl. Opt. **56**(16), 4675–4679 (2017).
27. C. Li, T. Ning, J. Li, L. Pei, C. Zhang, C. Zhang, H. Lin, and X. Wen, "Simultaneous measurement of refractive index, strain, and temperature based on a four-core fiber combined with a fiber Bragg grating," Opt. Laser Technol. **90**, 179–184 (2017).
28. G. Liu, M. Han, and W. Hou, "High-resolution and fast-response fiber-optic temperature sensor using silicon Fabry-Pérot cavity," Opt. Express **23**(6), 7237–7247 (2015).
29. Q. Sun, X. Sun, W. Jia, Z. Xu, H. Luo, D. Liu, and L. Zhang, "Graphene-Assisted Microfiber for Optical-Power-Based Temperature Sensor," IEEE Photonics Technol. Lett. **28**(4), 383–386 (2016).
30. H. Sun, S. Yang, X. Zhang, L. Yuan, Z. Yang, and M. Hu, "Simultaneous measurement of temperature and strain or temperature and curvature based on an optical fiber Mach–Zehnder interferometer," Opt. Commun. **340**, 39–43 (2015).
31. C. Ji, C.-L. Zhao, J. Kang, X. Dong, and S. Jin, "Multiplex and simultaneous measurement of displacement and temperature using tapered fiber and fiber Bragg grating," Rev. Sci. Instrum. **83**(5), 053109 (2012).

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/TIT.2019.2925613, IEEE Transactions on Information Theory

On Random Read Access in OCB

Ashwin Jha, Cuauhtemoc Mancillas-López, Mridul Nandi and Sourav Sen Gupta

***Abstract*—Offset Codebook or OCB mode is a popular block cipher mode of operation for authenticated encryption. The latest version of this cipher, called OCB3, is one of the finalists in CAESAR. In this paper we explore the scope of random read access and out-of-sequence decryption in OCB. We observe that the current versions of OCB are inefficient in this respect owing to the ineptness of the underlying mask generating function (MGF). We propose new candidates for MGF based on AES round function, which are efficient in direct computation and provide comparable performance in the usual setting. Our schemes are not the obvious choices for MGF in conventional sense as they do not have optimal almost XOR universal (AXU) bound. In existing OCB designs the MGFs are required to have 2^{-n}, i.e. optimal, AXU bound in order to upper bound the distinguishing advantage to $O(\sigma^2/2^n)$, where n is the block size of the underlying block cipher and σ is the total number of blocks among all queries. We find this specific requirement too restrictive. We abstract the OCB design, termed as GOCB, to look into the universal notion required from the underlying MGF. We propose a relaxed notion of AXU, called *locally imperfect XOR universal* (LIXU) hash, which can be of independent interest. Using LIXU as the underlying MGF we recover reasonable security bounds for our schemes.**

***Index Terms*—AES, AXU, masking, OCB, universal hash**

I. Introduction

Authenticated encryption or AE schemes [1], [2], [3] are symmetric-key algorithms which can simultaneously achieve data confidentiality, integrity and authentication. In recent years AE schemes have received considerable research interest, owing to the ongoing CAESAR competition [4] which aims to deliver a portfolio of state-of-the-art authenticated encryption schemes. Even though there are multiple designs based on pseudo-random generators [5], [6] and sponge-like functions [7], majority of the AE schemes have an underlying block cipher as their core component, and they rely on a secure and efficient mode of operation to achieve domain extension. In fact, 3 out of the 7 finalists in the CAESAR competition are based on an underlying (tweakable) block cipher. Some of the known block cipher based AE schemes are OCB [3], [8], [9], GCM [10], COLM [11], [12], [13] and Deoxys [14], [15].

Offset codebook or OCB mode [3], [8], [9] by Rogaway et al. is a single-pass and parallelizable block cipher mode of operation for nonce-based authenticated encryption (with associated data). OCB achieves both authentication and encryption with minimal performance overhead compared to classical single-pass encryption only modes such as CTR mode [16]. OCB requires one block cipher call per block of message, apart from a few constant number of calls for auxiliary tasks. In contrast, CCM [17] requires two block cipher calls per block of message, and GCM [10] requires one block cipher call and one field multiplication per block of message. As a result, OCB is approximately twice as fast [9], [18], [19] as any CTR mode based AE schemes, particularly CCM and GCM. In fact OCB3 [9] has, to a larger extent, already achieved the theoretical bounds for efficiency in software implementation.

Ashwin Jha and Mridul Nandi are with the Applied Statistics Unit, Indian Statistical Institute Kolkata, India (e-mail: ashwin.jha1991@gmail.com, mridul.nandi@gmail.com).

Cuauhtemoc Mancillas-López is with the Department of Computer Science, CINEVESTAV-IPN, Mexico (e-mail: cuauhtemoc.mancillas@cinvestav.mx).

Sourav Sen Gupta is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore (e-mail: sg.sourav@ntu.edu.sg).

A. Random Read Access and Out-of-Sequence Decryption

A block cipher based encryption mode of operation is said to have random read access feature, if it allows efficient decryption of any arbitrary encrypted block of message. This property could be beneficial in secure storage of read only data on cloud servers. The encrypted storage helps in avoiding unauthorized access, and the authentication tag helps in maintaining the integrity of the stored data. For example, say some enterprise application stores day to day system log files on some cloud server. Here the day timestamp may act as the nonce value. Now suppose on a later day the application faces some issue which warrants some portion of an encrypted log file. In this case the system administrator should be able to repeatedly read arbitrary blocks from the decrypted log, with as little overhead as possible, by simply mentioning the relevant day timestamp and the offset in the log file.

Yet another use-case for random read access is in disk encryption where access to an arbitrary block of some disk sector is highly beneficial. A popular and standardized disk encryption scheme, called XTS by Rogaway [8], is quite similar to OCB encryption.

A related feature is *out-of-sequence decryption where the mode of operation allows for the decryption of the encrypted block data stream in an arbitrary order*. In other words, the original ordering of the ciphertext blocks is not necessary for the decryption phase. If needed, the decrypted ciphertext can be reordered to get the original plaintext. This is quite useful in secure data transmission over connection-less network protocols such as User Datagram Protocol (UDP) [20], which allows for out-of-order delivery of data packets. If the underlying encryption scheme offers efficient out-of-sequence decryption then the data packets can be decrypted in on-the-fly manner; otherwise the received packets have to be reordered first which adds to the network protocol overhead. Yet another application area is real time data streaming protocols such as Secure Real-time Transport Protocol (SRTP) [21], [22], [23] and SRTP Control Protocol (SRTCP) [24], where on-the-fly decryption of received data packets is necessary to maintain

0018-9448 (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/TIT.2019.2925613, IEEE Transactions on Information Theory

the continuity of audio or visual data. In general, a scheme with random read access feature also allows out-of-sequence decryption.

Both CCM and GCM, being CTR based designs, allow for random read access: given the nonce and the block index, a single inverse block cipher call is required to access the message block. In case of OCB: given the nonce-based mask and the block index, a single block cipher call and a single MGF call is required to access the message block. So if the MGF is efficient in direct computation, then OCB can achieve random read access almost as efficiently as CCM and GCM.

We remark that random read access and out-of-sequence decryption are some features, which are defined exclusively for encryption schemes, and as such they allow release of unverified plaintext. It is well-known that OCB, CCM and GCM are trivially broken, when the adversary gets access to unverified plaintext. Still, we believe that there are practical use-cases, where random read access makes sense in context of authenticated encryption schemes as well. For instance, consider the use-case pertaining to encrypted log files on a cloud server. Here the cloud server may play the role of an adversary, in which case, it is clear that the adversary does not get access to unverified plaintext. Further, usually the logs follow some specific format for recording various events. If the encryption algorithm is secure, then it is almost impossible for the adversary to guess the right format. If the decryption algorithm receives unformatted log after decrypting few blocks, it can be sure that the cloud has tempered with the stored data.

B. Motivation

In connection with our discussion thus far, we observe a serious drawback in the instantiation of MGF in existing OCB designs. In all the existing OCB designs, the MGFs are instantiated with small-domain AXU hash functions. These functions are indeed quite efficient when computed sequentially over the indices; but they are not that efficient for direct computation at any arbitrary index, which renders the resulting OCB scheme inefficient in random read access. Take the use of xtimes in [3], [8], [9], [25] for example – the α-multiplication[1] is quite efficient in sequential computation as it requires only one shift and one conditional bit-wise XOR; but direct computation of xtimes for an arbitrary index i is not efficient at all, as it requires $O(\log i)$ many field multiplications if executed naively, or at least one field-multiplication even if all α^i values are precomputed, which is only possible for moderate values of i. Similar degradation is observed for other small-domain AXU hash functions such as Gray-code based hash [9] and Matrix-powered hash [26], [27]. In general, for most of the standard small-domain AXU hash functions: sequential computation requires just a few shift and XOR operations, whereas direct computation requires field multiplications or some other costly operations. This raises a simple question:

> (*a*) *Can we have efficient directly computable small-domain hash functions?*

[1]Here α denotes the primitive element of the underlying field.

A possible approach is to use the differential properties of data-dependent rotation (DDR) [28], [29], [9]. In fact OCB3 employs such a hash function for a very small domain of size 64, called stretch-then-shift [9], which requires four shifts and two XORs. Although DDR involves simple bit manipulation operations, the number of such operations are considerably higher to get a good AXU bound over a sufficiently large domain. For example the circulant hash [29] by Minematsu requires approx. 64 rotations and 64 XORs to get 2^{-128} AXU bound over $\mathbb{F}_{2^{64}}$. Furthermore, DDR is marred by patents, high cost in hardware and difficulty to get constant-time implementation.

Yet another approach is to use iterations of some cryptographic round function with good differential probability. Particularly, 4 rounds of AES offers very low differential probability bound [30], [31], [32] and most of the modern processors provide in-built support [33] for AES round function. This has lead to some efficient AES-based A(X)U hash functions [34], [35], [36]. While the use of reduced-round block ciphers could be an elegant solution, the number of rounds to get sufficient AXU bound is still a problem. For instance, on Intel Skylake processors, field multiplication is almost as good as four rounds of AES, which defeats the whole purpose of using AES rounds for MGF. For lesser number of rounds the performance is much better but the AXU bound is, in general, significantly low.

In view of this shortcoming of existing optimal AXU hash functions as MGFs, we rethink the basic philosophy behind masking in OCB like designs, and question the requirement of optimum AXU bounds in constructing the MGF. In existing proof techniques [8], [9], [37] the distinguishing advantage is upper bounded by $O(\sigma^2\epsilon)$, where σ is the total number of blocks in all queries and ϵ is the AXU bound of the MGF. Clearly, a sub-optimal ϵ will result in a sub-optimal bound for OCB. The immediate question that comes to mind is as follows:

> (*b*) *Can we relax the AXU condition on underlying MGFs to allow option for efficient direct computation while maintaining comparable security?*

In other words, we ask if one may reduce the AXU security margin of an underlying MGF without degrading the actual security of the OCB mode. More fundamentally, we try to investigate and establish the *practical requirements from a hash function* in terms of masking, rather than just accepting a standard off-the-shelf hash function with an optimum AXU bound from the literature.

C. Contributions

We answer the aforesaid questions (a) and (b) in affirmative, and hence our contributions are twofold:

- First, we propose two new 128-bit hash functions, called aes1 and aes2, based on 1 and 2 rounds of AES. We show that aes1 is a 2^{-96}-AXU hash over $\mathbb{F}_{2^8}$, and aes2 is a 2^{-113}-AXU hash over $\mathbb{F}_{2^{32}}$. On Intel Skylake processor: in direct computation, aes1 and aes2 are expected to be about 5 and 2.5 times faster than field multiplication. Similar improvements are expected for

0018-9448 (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/TIT.2019.2925613, IEEE Transactions on Information Theory

gray code and matrix based hash functions. In sequential computation aes1 and aes2 are slower than gray, but the degradation can be significantly reduced on modern processors where AES pipelining is possible.
Note that if we use previous proof techniques, we get bounds of $O(\sigma^2/2^{96})$ and $O(\sigma^2/2^{113})$ for OCB with aes1 and aes2, which are significantly worse then OCB3 bound, $O(\sigma^2/2^{128})$.

- Second, we introduce a generic view of OCB (termed GOCB), using the modular notion of Mask Generating Functions (MGFs), to study the alternatives to AXU hash functions for masking. We introduce a different notion of universal hash in the context of mask generation — the notion of *Locally-Imperfect XOR Universal Hash* (LIXU) — and prove that this notion is sufficient to provide adequate (comparable to existing) security guarantees for GOCB. The basic idea is to model a mask generating hash function H whose differential probability depends on grouping (quantified by the chunk number) of the given inputs. More precisely, for an (ϵ, r)-LIXU hash H the domain can be partitioned into several disjoint sets of size at most r, such that H has optimum differential probability when the inputs x and y belong to distinct partitions and the probability degrades to at most $\epsilon \geq 0$ when x and y belong to the same partition.
With respect to OCB, the standard notion of AXU uses a one-dimensional view of the message blocks, where the goal is to bound the differential probability between the masks of any two blocks among all σ many blocks. On the contrary, the notion of LIXU adopted by us views the message blocks conveniently as a two-dimensional array, where the differential probability is different within a row and across the rows. This view allows us to control the differential probability within a row by reducing the number of blocks (corresponds to the partition size r) in each message, while the differential probability across rows is controlled by the use of an encrypted nonce, as is the case with a regular AXU hash function. In fact, the number of *local* pairs of message blocks within a row may be controlled to be much less ($\leq r$) compared to the overall number of message blocks, such that even a low ϵ is adequately balanced in the effective security margin. This is the fundamental difference between an AXU hash and a LIXU hash, resulting in the benefits of the latter in terms of efficiency.
We obtain following security bound for GOCB, in Theorem 1, where the underlying MGF is an (ϵ, r)-LIXU hash function:

$$\mathbf{Adv}^{\mathsf{naead}}_{\mathsf{GOCB}[\Pi,\lambda]} \leq \frac{\sigma^2 + 3q\sigma + 2\sigma\widetilde{\sigma} + 2\widetilde{q}}{2^n} + (\sigma + 4\widetilde{\sigma})r\epsilon.$$

Based on this generalization, we recover the security bound of $O(\sigma^2 2^{-128})$ for GOCB based on aes1 and aes2 when the maximum message length is restricted to 2^6 and 2^{30} blocks, respectively. Note that, we take $r = 1$ when $\epsilon = \Theta(2^{-n})$, as this corresponds to a perfect AXU hash function.

STRUCTURE OF THE PAPER — In Section II, we define the preliminary notions and notation to aid the technical results, and Section III documents the existing constructions of small-domain AXU hash functions in the literature. In Section IV we propose our AES based hash functions and derive concrete AXU bounds for them. In Section V, we introduce the generic view of OCB, as well as the technical notion of LIXU hash functions. Section V also contains the main theorem on the security bound of GOCB using a LIXU hash function as MGF, and the concrete bounds for GOCB instantiation using aes1 and aes2. The security proofs are delegated to Section VI. Section VII gives a comparison between the performance of OCB3, and GOCB instantiated with aes1 and aes2 with respect to sequential and random access implementations. Finally, Section VIII concludes the paper.

II. PRELIMINARIES

NOTATION — In this paper, we abide by the following set of notation:

- $\mathbb{N}$ denotes the set of non-negative integers and $\mathbb{N}^+ = \mathbb{N} \setminus \{0\}$ the positives.
- For $n \in \mathbb{N}^+$, $[n]$ denotes the set $\{1, \ldots, n\}$.
- For $m, n \in \mathbb{N}$ and $m \leq n$, $[m..n]$ denotes the set $\{m, \ldots, n\}$.
- For $m \in \mathbb{N}^+$, I_m denotes the identity matrix of size m.
- For any set $\mathcal{S}$ and $m \in \mathbb{N}$, $\mathcal{S}^m$ denotes the set of all m-tuples from $\mathcal{S}$, and $\mathcal{S}^{\underline{m}}$ denotes the set of all m-tuples with distinct elements from $\mathcal{S}$.
- We let $\mathcal{S}^{\leq \ell} = \cup_{i=1}^{\ell} \mathcal{S}^i$, $\mathcal{S}^+ = \cup_{i=1}^{\infty} \mathcal{S}^i$, and $\mathcal{S}^* = \cup_{i=0}^{\infty} \mathcal{S}^i$.

Throughout this paper, we fix $n \in \mathbb{N}^+$ as the block size, and the elements of $\{0,1\}$ and $\{0,1\}^n$ (often denoted by $\mathcal{B}$) are called *bits* and *blocks*, respectively. We use $\oplus$ and $\odot$ to denote the field addition (XOR) and field multiplication, respectively, over the finite field $\mathbb{F}_{2^n} = \mathcal{B} = \{0,1\}^n$. Throughout the paper, α denotes a primitive element of $\mathbb{F}_{2^n}$, and for any $x \in \mathbb{F}_{2^n}$, the operation $x \odot \alpha$ will be referred to as the α-multiplication operation on x.

REPRESENTATION OF STRINGS — If x is a vector (or a sequence or string) over a finite set $\mathcal{S}$ (coordinates are from $\mathcal{S}$), then $|x|$ typically denotes its length. For $i \in [|x|]$, x_i denotes its i-th coordinate.

- For $i, j \in [|x|]$ and $i \leq j$, we use the shorthand $x_{i..j} = (x_i, \ldots, x_j)$.
- For two strings x and y, $x||y$ denotes the concatenation of x and y.
- If $z = x\|y$, then x and y are the *prefix* and *suffix*, respectively of z.
- Strings over $\{0,1\}$ and $\{0,1\}^n$ are called binary and block strings respectively, and we will only consider binary and block strings in this paper.
- For a string b and $m \in \mathbb{N}$, b^m denotes the m-wise concatenation of b, i.e $b\|b\|\cdots$ (m times), this being the null string when $m = 0$.
- For some $m \in \mathbb{N}^+$, for $x \in \{0,1\}^m, b \in [m]$, $x \gg b$ denotes x shifted b bits to the right, i.e., $0^b\|x_{1..m-b}$, while $x \ll b$ denotes x shifted b bits to the left.

0018-9448 (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/TIT.2019.2925613, IEEE Transactions on Information Theory

NOMENCLATURE FOR BLOCKS — For a binary string x, if $|x| = n$, we call x a complete block; if $|x| < n$, we call x an incomplete block; if $|x| = 0$, we call x the null string or the empty block. By convention, the empty block will also be referred as an incomplete block. Any binary sequence x can be mapped uniquely to a block sequence $\hat{x} = (\hat{x}_1, \ldots, \hat{x}_\ell, \hat{x}_*) \xleftarrow{n} x$, where $\ell \in \mathbb{N}$, $\hat{x}_1, \ldots, \hat{x}_\ell$ are complete blocks, and $\hat{x}_*$ is an incomplete (possibly empty) block, such that $\hat{x} = \hat{x}_1 \| \ldots \| \hat{x}_\ell \| \hat{x}_*$. For this mapping we take $\ell = \lfloor |x|/n \rfloor$, $\hat{x}_i = x_{n(i-1)+1..ni}$ for $i \in [\ell]$, and $\hat{x}_* = x_{n\ell+1..|x|}$. For an incomplete block x, $\mathsf{pad}(x)$ denotes the complete block $x\|10^* = x\|1\|0^{n-|x|-1}$. For a complete block x and $k \in [n]$, $\mathsf{msb}_k(x)$ and $\mathsf{lsb}_k(x)$ denote the substring $x_{1..k}$ and $x_{n-k+1..n}$ respectively. For a binary string x, $\mathsf{ntz}(x)$ denotes the number of trailing zero bits in x.

RANDOM FUNCTIONS AND PERMUTATIONS — For a set $\mathcal{S}$, we write $x \leftarrow\!\!\$\, \mathcal{S}$ to denote that x is sampled from $\mathcal{S}$ uniformly. For a given domain $\mathcal{D}$ and a given co-domain $\mathcal{R}$, we denote by $\mathsf{Func}[\mathcal{D}, \mathcal{R}]$ the set of all functions from $\mathcal{D}$ into $\mathcal{R}$. We say Γ is an ideal random function from $\mathcal{D}$ to $\mathcal{R}$ to indicate that $\Gamma \leftarrow\!\!\$\, \mathsf{Func}[\mathcal{D}, \mathcal{R}]$. On a similar note, we denote by $\mathsf{Perm}[\mathcal{D}]$ the set of all permutations on $\mathcal{D}$. We say Π is an ideal random permutation on $\mathcal{D}$ to indicate that $\Pi \leftarrow\!\!\$\, \mathsf{Perm}[\mathcal{D}]$. For distinct inputs $x_1, \ldots, x_m \in \mathcal{D}$, $(\Gamma(x_1), \ldots, \Gamma(x_m)) \leftarrow\!\!\$\, \mathcal{R}^m$, and $(\Pi(x_1), \ldots, \Pi(x_m)) \leftarrow\!\!\$\, \mathcal{D}^{\underline{m}}$. In other words, an ideal random function Γ can be viewed as a *with-replacement* sampler from $\mathcal{R}$, whereas an ideal random permutation Π can be viewed as a *without-replacement* sampler from $\mathcal{D}$.

A. Distinguishers and Advantage

Given two oracles $\mathcal{O}_0$ and $\mathcal{O}_1$, an algorithm $\mathscr{A}$ that tries to distinguish between $\mathcal{O}_0$ and $\mathcal{O}_1$ is called a *distinguishing adversary* or a *distinguisher*. $\mathscr{A}$ plays an interactive game with $\mathcal{O}_b$ (for some bit b) in a black box fashion, and outputs a single bit at the end. For an oracle $\mathcal{O}$, $\mathscr{A}^{\mathcal{O}}$ denotes $\mathscr{A}$'s output after its interaction with $\mathcal{O}$, and we say that $\mathscr{A}$ wins the distinguishing game if $\mathscr{A}^{\mathcal{O}} = b$. The distinguishing advantage of $\mathscr{A}$ is defined as

$$\mathbf{Adv}_{\mathcal{O}_1;\mathcal{O}_0}(\mathscr{A}) := \left| \Pr\left[\mathscr{A}^{\mathcal{O}_1} = 1\right] - \Pr\left[\mathscr{A}^{\mathcal{O}_0} = 1\right] \right|.$$

Let $\mathbb{A}(q, \ell, \sigma, t)$ be the class of all distinguisher limited to q oracle queries, each of length at most ℓ blocks, total length $\sigma \le q\ell$, and t computations. We define

$$\mathbf{Adv}_{\mathcal{O}_1;\mathcal{O}_0}(q, \ell, \sigma, t) := \max_{\mathscr{A} \in \mathbb{A}(q,\ell,\sigma,t)} \mathbf{Adv}_{\mathcal{O}_1;\mathcal{O}_0}(\mathscr{A}).$$

NOTIONS OF VARIOUS ORACLES — $\mathcal{O}_0$ conventionally represents an ideal primitive, while $\mathcal{O}_1$ represents either an actual construction or a mode of operation built of some other ideal primitives. Typically the goal of the function represented by $\mathcal{O}_1$ is to emulate the ideal primitive represented by $\mathcal{O}_0$. We use the standard terms *ideal oracle* and *real oracle* for $\mathcal{O}_0$ and $\mathcal{O}_1$ respectively.

SECURITY GAME — In the context of this paper, a security game is a standard distinguishing game with an optional set of additional restrictions, chosen to reflect the desired security goal. For instance, some security games give bidirectional access to the underlying function of an oracle $\mathcal{O}$ and its inverse. This is denoted by $\mathcal{O}^{\pm}$. When we talk of distinguishing advantage with a specific security game G in mind, we include G in the superscript, e.g., $\mathbf{Adv}^{G}_{\mathcal{O}_1;\mathcal{O}_0}(\mathscr{A})$. Usually, $\mathcal{O}_0$ is completely identified by G, and hence dropped from the subscript.

B. PRF and PRP Security Games

Let $f : \mathcal{K} \times \mathcal{D} \to \mathcal{R}$ be a keyed function family from $\mathcal{D}$ to $\mathcal{R}$ indexed by the key space $\mathcal{K}$. In the pseudorandom function security game prf, the goal of any adversary $\mathscr{A}$ is to distinguish f from an ideal random function $\Gamma \leftarrow\!\!\$\, \mathsf{Func}[\mathcal{D}, \mathcal{R}]$. Hence the PRF-advantage of $\mathscr{A}$ against f is defined as

$$\mathbf{Adv}^{\mathsf{prf}}_{f}(\mathscr{A}) := \left| \Pr_{K \leftarrow\!\!\$\, \mathcal{K}}\left[\mathscr{A}^{f_K} = 1\right] - \Pr_{\Gamma}\left[\mathscr{A}^{\Gamma} = 1\right] \right|.$$

With q, ℓ, σ and t defined as above, we define the PRF-security of f against the class of adversaries $\mathbb{A}(q, \ell, \sigma, t)$ as $\mathbf{Adv}^{\mathsf{prf}}_{f}(q, \ell, \sigma, t)$. In the same way we can define PRP or pseudorandom permutation advantage. For a keyed permutation f over $\mathcal{R}$, we define

$$\mathbf{Adv}^{\mathsf{prp}}_{f}(\mathscr{A}) := \left| \Pr_{K \leftarrow\!\!\$\, \mathcal{K}}\left[\mathscr{A}^{f_K} = 1\right] - \Pr_{\Gamma}\left[\mathscr{A}^{\Pi} = 1\right] \right|.$$

In a stronger variant of PRP game, called strong PRP (SPRP) security game, the adversary is given access to both the forward and inverse queries to the oracle. The SPRP advantage of $\mathscr{A}$ is defined analogously as $\mathbf{Adv}^{\mathsf{sprp}}_{f}(\mathscr{A})$.

C. Authenticated Encryption Security Game

A *nonce-based authenticated encryption scheme with associated data* (NAEAD) consists of a key space $\mathcal{K}$, a message space $\mathcal{M}$, an associated data space $\mathcal{A}$, a nonce space $\mathcal{N}$ and a tag space $\mathcal{T}$, along with two functions $\mathsf{Enc} : \mathcal{K} \times \mathcal{N} \times \mathcal{A} \times \mathcal{M} \to \mathcal{M} \times \mathcal{T}$ and $\mathsf{Dec} : \mathcal{K} \times \mathcal{N} \times \mathcal{A} \times \mathcal{M} \times \mathcal{T} \to \mathcal{M} \cup \{\perp\}$, with the correctness condition that for any $K \in \mathcal{K}, N \in \mathcal{N}, A \in \mathcal{A}, M \in \mathcal{M}$, we have

$$\mathsf{Dec}(K, N, A, \mathsf{Enc}(K, N, A, M)) = M.$$

In addition, in most popular authenticated encryption schemes (including $\mathsf{OCB3}$), the map $\mathsf{proj}_{\mathcal{M}} \circ \mathsf{Enc}(K, N, A, \cdot)$ for fixed K, N, A is a length-preserving permutation, where $\mathsf{proj}_{\mathcal{M}} : \mathcal{M} \times \mathcal{T} \to \mathcal{M}$ is the projection on $\mathcal{M}$.

In the nonce-respecting authenticated encryption security game with associated data naead, $\mathcal{O}_1^+$ and $\mathcal{O}_1^-$ of the real oracle are the encryption function $\mathsf{Enc}(K, \cdot, \cdot, \cdot)$ and decryption function $\mathsf{Dec}(K, \cdot, \cdot, \cdot, \cdot)$ respectively of the authenticated encryption scheme under consideration for a key K randomly chosen from $\mathcal{K}$; in the ideal oracle, $\mathcal{O}_0^+$ and $\mathcal{O}_0^-$ are $\Gamma \leftarrow\!\!\$\, \mathsf{Func}[\mathcal{N} \times \mathcal{A} \times \mathcal{M}, \mathcal{M} \times \mathcal{T}]$, and $\perp : \mathcal{N} \times \mathcal{A} \times \mathcal{M} \times \mathcal{T} \to \{\perp\}$ respectively. We henceforth refer to $(\Gamma, \perp)$ as the *ideal nonce-based authenticated encryption scheme*. The distinguishing adversary operates under the following restrictions — (a) no two encryption queries can have the same

0018-9448 (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/TIT.2019.2925613, IEEE Transactions on Information Theory

nonce, and (b) if an encryption query (N, A, M) yields (C, T), a decryption query (N, A, C, T) is not allowed.

SECURITY GOALS — Distinguishing advantage of any adversary $\mathscr{A}$ in naead with associated data will be denoted as $\mathbf{Adv}^{\mathsf{naead}}_{\mathcal{O}_1;\mathcal{O}_0}(\mathscr{A})$. Note that security under this formulation covers the two standard security goals of authenticated encryption:

- *Privacy* — Security against an adversary who tries to distinguish the AE or AEAD construction from an ideal random function Γ, and
- *Integrity* — Security against an adversary who tries to make a successful forging attempt on the scheme using decryption queries.

III. REVISITING SMALL-DOMAIN AXU HASH FUNCTIONS

An $(\mathcal{L}, \mathcal{D}, \mathcal{R})$-hash function $H : \mathcal{L} \times \mathcal{D} \to \mathcal{R}$ is a keyed function family from *domain or message space* $\mathcal{D}$ to *digest or hash space* $\mathcal{R}$ and indexed by the *key space* $\mathcal{L}$. In this work we fix $\mathcal{R} = \mathcal{B}$ and hence drop it from the notation.

Definition 1 (AXU hash function [38], [39])**.** An $(\mathcal{L}, \mathcal{D})$-hash function H is called ϵ-*AXU* (almost XOR universal) hash function if $\forall\ X \neq Y \in \mathcal{D}$, $\delta \in \mathcal{B}$,

$$\mathsf{DP}_H(X, Y : \delta) := \Pr_{L \xleftarrow{\$} \mathcal{L}}[H_L(X) \oplus H_L(Y) = \delta] \leq \epsilon. \quad (1)$$

The term $\mathsf{DP}_H(X, Y : \delta)$ is also called the differential probability (or DP) of H on inputs X and Y and difference δ. The above definition says that the maximum differential probability (or MDP) of H is at most ϵ.

COMPOSITIONS — We consider two compositions of AXU hash functions:

- *Range Extension* — For ϵ_i-AXU hash functions $H_i : \mathcal{L}_i \times \mathcal{D} \to \mathcal{R}_i$, $i = 1, 2$,

$$H_1 \| H_2 : (\mathcal{L}_1 \times \mathcal{L}_2) \times \mathcal{D} \to \mathcal{R}_1 \times \mathcal{R}_2$$
$$(L_1, L_2, X) \mapsto H_1(L_1, X) \| H_2(L_2, X)$$

- *Domain Extension* — For ϵ_i-AXU hash functions $H_i : \mathcal{L}_i \times \mathcal{D}_i \to \mathcal{R}$, $i = 1, 2$,

$$H_1 \oplus H_2 : (\mathcal{L}_1 \times \mathcal{L}_2) \times (\mathcal{D}_1 \times \mathcal{D}_2) \to \mathcal{R}$$
$$((L_1, L_2), (X_1, X_2)) \mapsto H_1(L_1, X_1) \oplus H_2(L_2, X_2)$$

It is easy to verify that in range extension, $H_1 \| H_2$ is an $\epsilon_1\epsilon_2$-AXU hash function if L_1 and L_2 are independently sampled. It can also be verified that the domain extension hash $H_1 \oplus H_2$ is a $\max\{\epsilon_1, \epsilon_2\}$-AXU hash function if L_1 is independent of L_2. We summarize the two observations in the following proposition.

Proposition 1. *Let H_i be ϵ_i-AXU hash functions for $i \in \{1, 2\}$. Then,*

- *$H_1 \| H_2$ hash function is an $\epsilon_1\epsilon_2$-AXU hash function; and*
- *$H_1 \oplus H_2$ hash function is a $\max\{\epsilon_1, \epsilon_2\}$-AXU hash function.*

A. Examples of Small-Domain AXU Hash Function

Symmetric-key cryptographic literature is rich in AXU hash function candidates. In this work we concentrate only on *small-domain* hash functions that can be used to generate masks in OCB. In what follows, we discuss some examples of AXU hash functions, with a summary as in Table I.

TABLE I: Feature summary of some Small-Domain AXU Hash Function candidates, where $\oplus$, $\odot$, $\ll$, $\|$, $\oplus_?$ denote addition, multiplication, shifts, concatenation and conditional addition over appropriate fields; AESRD denotes AES round function, wt and ntz denotes the hamming weight and trailing zeros function.

Hash Family	Sequential (operations)	Direct (operations)	Key Size	Domain Size in log base 2	MDP
xtimes [3]	$1 \ll, 1 \oplus_?$	$O(\lg i) \odot$	128	128	2^{-128}
gray [25], [9]	1 ntz(i), $1 \oplus$	$1 \oplus, 1 \ll, 1 \odot$	128	128	2^{-128}
mtrx [26], [27]	$2 \ll, 1 \oplus$ [#]	$O(\lg i) \odot$	128	128	2^{-128}
mlin	wt$(i) \oplus$	wt$(i) \oplus$	$128r$	r	2^{-128}
clh [29]	wt$(i) \ll$, wt$(i) \oplus$	wt$(i) \ll$, wt$(i) \oplus$	128	127	2^{-128}
sts [9]	$1 \|, 1 \oplus, 3 \ll$	$1 \|, 1 \oplus, 3 \ll$	128	6	2^{-128}
lcube	$1 \oplus, 2 \odot$	$1 \oplus, 2 \odot$	128	128	2^{-128}
linv [3]	$1 \oplus, 1 \odot$	$1 \oplus, 1 \odot$	128	128	2^{-128}
aes4 [35]	4 AESRD	4 AESRD	512	128	2^{-113}
aes2 (Section IV)	2 AESRD	2 AESRD	256	32	2^{-113}
aes1 (Section IV)	1 AESRD	1 AESRD	128	8	2^{-96}

[#] Note that [26] defines many three-operation maskings for 128 bit binary field.

Example 1 (α-multiplication based hash [3])**.** For some reasonably large $\ell \in [2^n - 1]$, let $\mathcal{D} = [\ell]$ and $\mathcal{L} = \mathcal{B}$. The hash function xtimes, defined as $\forall L \in \mathcal{B}, i \in [\ell]$,

$$\mathsf{xtimes}_L(i) = \mathsf{xtimes}(L, i) := \alpha^i \odot L, \quad (2)$$

is a 2^{-n}-AXU hash. xtimes has been used to generate position dependent input masks in many symmetric-key designs including OCB designs [3], [8], [9], OTR [40], PMAC [25], PMAC_Plus [41], EME [42], COPA [11] and ELmD [12]. The popularity of xtimes is mainly due to its efficiency in sequential computation, as we have

$$\forall i \geq 1,\ L \in \mathcal{B},\ \mathsf{xtimes}_L(i+1) = \alpha \odot \mathsf{xtimes}_L(i),$$

and α-multiplication is efficient (one shift and one conditional bit-wise XOR operation). While the sequential computation of $\mathsf{xtimes}_L(i)$ is very efficient, the same is not reflected in direct computation of $\mathsf{xtimes}_L(i)$ when i is a large integer. In general it may require $O(\lg i)$ many field multiplications. Even when the α^i values are precomputed, we need one field multiplication which is not as efficient as one α-multiplication. Indeed we can even avoid the precomputation and simply use $i \odot L$ as the hash definition.

Example 2 (Gray-code based hash [25], [9])**.** For reasonably large $\ell \in [2^n - 1]$, let $\mathcal{D} = [\ell]$ and $\mathcal{L} = \mathcal{B}$. The hash function gray is defined as,

$$\forall L \in \mathcal{B},\ i \in [\ell],\quad \mathsf{gray}_L(i) = \mathsf{gray}(L, i) := \gamma(i) \odot L \quad (3)$$

where $\gamma(i) := i \oplus (i \gg 1)$ is the gray code value for i. The Gray code sequence has a very nice feature that for $i \geq 1$, the i-th gray code can be written in terms of the $(i-1)$-th gray code, i.e

$$\gamma(i) = \gamma(i-1) \oplus 2^{\mathsf{ntz}(i)},$$

0018-9448 (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/TIT.2019.2925613, IEEE Transactions on Information Theory

When we substitute this in (3) we get an alternate definition $\forall L \in \mathcal{B},\ i \in [\ell]$,

$$\mathsf{gray}_L(i) = \mathsf{gray}(L, i) := \oplus_{j=1}^{i} L_{\mathsf{ntz}(j)}. \quad (4)$$

where $L_j = \alpha^j \odot L$ for all $j < n$. In [9] and RFC7253 [43], gray has been shown to be a 2^{-n} AXU hash function. Note that we have

$$\forall i \geq 1,\ L \in \mathcal{B},\ \mathsf{gray}_L(i+1) = \mathsf{gray}_L(i) \oplus L_{\mathsf{ntz}(i+1)}.$$

So if the L_j values are precomputed, then in sequential computation gray hash only requires, ntz computation along with a single n-bit XOR operation. Hence it is very efficient in sequential computations. This hash function is applied in OCB3 [9] (RFC7253 [43]) and PMAC [25].

Example 3 (Matrix-powered hash [26], [27])**.** Let $n = wn'$ for some positive integer w, called word size, whose values are generally software friendly such as 8, 16, 32, 64 etc. Let $\mathcal{W} = \{0,1\}^w$ and $\mathcal{L} = \mathcal{W}^{n'}$ and $\mathcal{D} = [\ell]$ for some reasonably large $\ell \in [2^n - 1]$. Let M be an invertible $n' \times n'$ matrix whose entries are from the field[2] $\mathcal{W}$ such that for all $i \in [2^n - 1]$, $I_{n'} + M^i$ is invertible. Then the hash function mtrx (called *matrix-powered hash*), defined as

$$\forall L \in \mathcal{W}^{n'}, i \in [\ell],\quad \mathsf{mtrx}_L(i) = \mathsf{mtrx}(L, i) := M^i \cdot L, \quad (5)$$

is a 2^{-n}-AXU. Note here that L is viewed as a vector over $\mathcal{W}$. Various matrix-powered hash candidates are given in [26] and [27]. But, similar to xtimes hash, all those hash functions are efficient in sequential computation and inefficient in direct computation.

The matrix-powered hash function is a very general candidate. For example, xtimes and gray hash functions can be viewed as instances of matrix-powered hash for appropriate choices of the matrix M. The readers are directed to [26] and [27] for a detailed exposition on matrix-powered hash functions. A table of various word-oriented matrices has been listed in [26].

Example 4. Some other finite field based constructions, with $\mathcal{L} = \mathcal{D} = \mathcal{B}$, are the following: $\forall L \in \mathcal{B},\ x \in \mathcal{B}$

- $\mathsf{lcube}_L(x) = (L \oplus x)^3$;
- $\mathsf{linv}_L(x) = (L \oplus x)^{-1}$ (if $x \neq L$), 0 otherwise.

It is easy to verify that lcube is 2^{1-n}-AXU and linv is 2^{2-n}-AXU hash function.

Example 5 (Multi-linear hash)**.** For $\ell \in [2^n - 1]$, let $\mathcal{D} = [0..\ell]$ and $\mathcal{L} = \mathcal{B}^{\lg \ell}$. The hash function mlin (called multi-linear hash), is defined as

$$\forall L \in \mathcal{L},\ b \in \mathcal{D},\quad \mathsf{mlin}_L(b) = \mathsf{mlin}(L, b) := \oplus_{i=1}^{r} b_i L_i, \quad (6)$$

is a 2^{-n}-AXU hash function, where $L_1, \ldots, L_r \in \mathcal{B}$ and $b_1, \ldots, b_r \in \{0,1\}$. In Grain-128a [44], the tag is computed using mlin, where the keys are derived using Toeplitz matrix.

Example 6 (Circulant hash [29])**.** In [29] Minematsu presented a small-domain hash function based on data-dependent rotations. For $\ell \in [2^{n-1} - 1]$, let $\mathcal{D} = [0..\ell]$ and $\mathcal{L} = \mathcal{B}$. The hash function clh (called circulant hash) is defined as $\forall K \in \mathcal{L},\ x \in \mathcal{D}$,

$$\mathsf{clh}_L(x) = \mathsf{clh}(L, x) := \bigoplus_{i \in [\lg \ell]:\ x_i = 1} (L \ll (i-1)). \quad (7)$$

For direct computation, the hash function might require $n-1$ rotations and $n-2$ XORs in the worst case; for sequential computation some optimization is possible for the i-th input given the $(i-1)$-th output. In [29], clh is shown to be both 2^{-128}-AXU and 2^{-128}-uniform hash function.

Example 7 (Stretch-then-shift hash [9])**.** In [9] Krovetz and Rogaway presented a highly efficient AXU hash function over a small domain $\mathcal{D} = [64]$. It follows the general data-dependent rotation technique by Minematsu [29]. Fix $c \in \mathbb{N}^+$. For $\mathcal{D} = [0..\psi(c) - 1]$ and $\mathcal{L} = \mathcal{B}$, where $\psi(c)$ is a positive integer denoting the maximum domain size for a fixed c. The hash function sts (called stretch-then-shift hash) is defined as $\forall L \in \mathcal{B},\ i \in [0..\psi(c) - 1]$,

$$\mathsf{sts}_L(i) = \mathsf{sts}(L, i) := \mathsf{msb}_n(\mathsf{stretch}_c(L) \ll i), \quad (8)$$

where $\forall L \in \mathcal{B},\ \ \mathsf{stretch}_c(L) := L \| (L \oplus (L \ll c))$. The hash function first stretches the key based on a parameter c, and then makes data-dependent shift before outputting the most significant n bits; hence the name. For $n = 128$, $c = 8$ and $\psi(c) = 64$, Krovetz and Rogaway showed that sts achieves 2^{-128}-AXU and 2^{-128}-uniform bound [9].

IV. AES-BASED SMALL-DOMAIN AXU HASH FUNCTIONS

In the previous section we listed some hash functions based on finite field operations and data dependent rotations. Now we will propose some hash functions based on the differential properties of the AES rounds. Before moving forward, we give the standard and well-known [35], [30] definition for differential probability of a (keyed) permutation.

Definition 2. Let π be a permutation over $\mathbb{F}_{2^n}$ and Π_K be a keyed permutation over $\mathbb{F}_{2^n}$ with key $K \leftarrow\!\!\$\, \mathcal{K}$. For any given non-zero $a, b \in \mathbb{F}_{2^n}$ the differential probability of π, is defined as

$$\mathsf{dp}_\pi(a, b) := \Pr_X[\pi(X) \oplus \pi(X \oplus a) = b].$$

The maximum differential probability (MDP) of π is defined as

$$\mathsf{mdp}(\pi) := \max_{a \neq 0, b} \mathsf{dp}_\pi(a, b).$$

Similarly, the expected differential probability (EDP) of Π_K, is defined as

$$\begin{aligned}\mathsf{edp}_{\Pi_K}(a, b) &:= \Pr_{K,X}[\Pi_K(X) \oplus \Pi_K(X \oplus a) = b] \\ &= \sum_{K \in \mathcal{K}} \Pr_X[\Pi_k(X) \oplus \Pi_k(X \oplus a) = b \mid k] \cdot \Pr[K = k] \\ &= \sum_{K \in \mathcal{K}} \mathsf{dp}_{\Pi_k}(a, b) \cdot \Pr[K = k] = \mathbb{E}_K[\mathsf{dp}_{\Pi_K}(a, b)],\end{aligned}$$

[2] $\{0,1\}^m$ can be viewed as the binary field by fixing a $\deg m$ primitive polynomial.

0018-9448 (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/TIT.2019.2925613, IEEE Transactions on Information Theory

where $\mathbb{E}_K$ denotes the expectation over K. The maximum expected differential probability (MEDP) of Π_K is defined as

$$\mathsf{medp}(\Pi_K) := \max_{a \neq 0, b} \mathsf{edp}_{\Pi_K}(a, b).$$

Note that we have also used the term differential probability in the context of AXU hash functions in Definition 1. Now that we have formally defined the differential probability of a (keyed) permutation, it is imperative that we discuss the similarity and differences between the two probabilities. The major difference is the source of randomness. In the AXU view of differential probability, the randomness is due to the hash key, whereas here one of the input is chosen randomly. Even though this seems to be a major difference, but for certain class of functions the two views have a very nice relationship. Suppose π is a (possibly keyed) permutation over $\mathbb{F}_{2^n}$ with keyspace $\mathcal{K}$ (empty set when π is keyless). We define a function $\pi' : \mathbb{F}_{2^n} \times \mathcal{K} \times \mathbb{F}_{2^n} \to \mathbb{F}_{2^n}$ as a keyed family of function indexed by $K, K' \in \mathbb{F}_{2^n} \times \mathcal{K}$, such that

$$\forall K, K', x \in \mathbb{F}_{2^n} \times \mathcal{K} \times \mathbb{F}_{2^n}, \quad \pi'_{K,K'}(x) := \pi_{K'}(K \oplus x).$$

Lemma 1 is a well-known result that we prove here just for the sake of completeness.

Lemma 1. *For some $\epsilon \geq 0$,*

1) *If π is keyless and $\mathsf{mdp}(\pi) = \epsilon$ then π' is an ϵ-AXU hash for $K \xleftarrow{\$} \mathbb{F}_{2^n}$.*
2) *If π is keyed and $\mathsf{medp}(\pi) = \epsilon$ then π' is an ϵ-AXU hash for $K \xleftarrow{\$} \mathbb{F}_{2^n}$ and $K' \xleftarrow{\$} \mathcal{K}$.*

Proof. We prove the unkeyed version of the lemma. The keyed version can be proved in similar fashion. Let $x, x' \in \mathbb{F}_{2^n}$ such that $x \neq x'$ and $\delta \in \mathbb{F}_{2^n}$. If $\delta = 0$ then the result is vacuously true as π' is a permutation. Suppose $\delta \neq 0$. Then, we have

$$\begin{aligned}\mathsf{DP}_{\pi'}(x, x' \ : \ \delta) &= \Pr_K[\pi'(K \oplus x) \oplus \pi'(K \oplus x') = \delta] \\ &= \Pr_X[\pi'(X) \oplus \pi'(X \oplus x' \oplus x) = \delta] \\ &\leq \mathsf{dp}(\pi) = \epsilon.\end{aligned}$$

□

A. Revisiting MEDP bounds for the AES

The AES block cipher is a Substitution Permutation Network (SPN) with block size n = 128, internal s-box input size b = 8, where all s-boxes are identical, denoted by S. The permutation layer, denoted by ϕ, consists of a bytewise permutation followed by four identical 32-bit linear transformations applied in parallel. One round of AES, denoted by 1-AESRD (without the subkey mixing) is nothing but $\phi \circ S$. Using the 1-AESRD round function, we can define the keyed function i-AESRD for all integers $i > 1$ as follows: $\forall x \in \mathbb{F}_{2^{128}}, K \in \mathbb{F}_{2^{128}}^{i-1}$,
For $i = 2$:

$$2\text{-AESRD}_K(x) = 1\text{-AESRD}(K_1 \oplus 1\text{-AESRD}(x))$$

For $i > 2$:

$$i\text{-AESRD}_K(x) = 1\text{-AESRD}(K_{i-1} \oplus (i-1)\text{-AESRD}_{K_{[i-2]}}(x))$$

The differential properties of AES has been a well-studied [45], [30], [31], [32], [46] topic in symmetric-key cryptography. In fact, one of the design criteria for AES [45], [47] was protection against linear [48] and differential cryptanalysis [49].

In [46], Daemen and Rijmen first observed that due to the specific nature of the AES permutation layer ϕ, the differential probabilities over 2 AES rounds are equivalent to those over a reduced SPN structure, which they called a *super box*. Basically the two rounds of AES over 128-bit input can be viewed as four parallel invocations of the super box with independent keys (disjoint substrings of a uniform string are independent) each working with distinct 32 bits of the input. We denote an AES super box instantiated with a subkey K by $\mathtt{aesSB}_K$.

Keliher and Sui [31], [32] later used this idea to derive a tight MEDP bound for 2-AESRD. Specifically, in [31, Theorem 1, Theorem 2] and [32, section 4.1] they proved that the MEDP for AES super box is 1.656×2^{-29}, which is equivalent to the MEDP for 2-AESRD. We refer the readers to [46], [32] for further exposition on AES super box and its relation with 2-AESRD. Using the fact that the upper bound on the MEDP for 4 or more rounds of AES is equal to the 4^{th} power of the upper bound on the MEDP for 2-AESRD, Keliher and Sui [32] further showed that MEDP for t-AESRD for $t \geq 4$ is upper bounded by 1.881×2^{-114}. We summarize these results in Proposition 2.

Proposition 2. *[32, section 4.1] Let X_1, X_2, X_3 be uniformly and independently sampled from $\mathbb{F}_{2^{128}}$ and $Y \xleftarrow{\$} \mathbb{F}_{2^{32}}$. Then,*

1) $\mathsf{medp}(\mathtt{aesSB}_Y) \approx 1.656 \times 2^{-29}$.
2) $\mathsf{medp}(2\text{-AESRD}_{X_1}) \approx 1.656 \times 2^{-29}$.
3) $\mathsf{medp}(4\text{-AESRD}_{X_1,X_2,X_3}) \leq 1.881 \times 2^{-114}$.

B. Our AES-based proposals

In the following discussion we fix $n = 128$, i.e. $\mathcal{B} = \{0,1\}^{128}$. Before presenting our proposals we start with a simple 4-AESRD based hash function in the following example.

Example 8. Let $\mathcal{D} = \mathcal{B}$ and $\mathcal{L} = \mathcal{B}^4$. For $L = (L_1, L_2, L_3, L_4) \in \mathcal{B}^4$ and $x \in \mathcal{B}$, the 4-AESRD based hash function aes4 is defined as

$$\mathsf{aes4}_L(x) = \mathsf{aes4}(L, x) := 4\text{-AESRD}_{L_2,L_3,L_4}(L_1 \oplus x).$$

From Lemma 1 and Proposition 2, it is straightforward to see that aes4 is a 1.88×2^{-114}-AXU hash function. In [35] Minematsu and Tsunoo used a variant of aes4 to construct a universal hash function and subsequently extended it to construct a MAC. Jakimoski and Subbalakshmi [50] further built upon their work to get more efficient hash functions and MAC.

1) 1-Round AES based Hash Function.: The MDP for AES s-box S is known to be 2^{-6} [30]. We define a keyed function $S' : \mathbb{F}_{2^8} \times \mathbb{F}_{2^8} \to \mathbb{F}_{2^8}$ as

$$\forall K, x \in \mathbb{F}_{2^8}^2, \quad S'_K(x) := S(K \oplus x).$$

Clearly S and S' satisfy the conditions in Lemma 1, hence S' is 2^{-6}-AXU hash function. Now, let $\mathcal{D} = \{0,1\}^8$ (viewed as

0018-9448 (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/TIT.2019.2925613, IEEE Transactions on Information Theory

$[256]$) and $\mathcal{L} = \mathcal{B}$ (viewed as $\mathcal{D}^{16}$). Given a key L, and an input $i \in \mathcal{D}$, we define the one round AES hash, aes1 as

$$\begin{aligned}\mathsf{aes1}_L(i) = \mathsf{aes1}(L, i) &:= \mathsf{1\text{-}AESRD}(L \oplus i^{16}) \\ &= \phi(S'_{L_1}(i) \| \cdots \| S'_{L_{16}}(i)). \end{aligned} \quad (9)$$

Recall that ϕ is the linear transformation or permutation layer used in AES. It is well-known that the AXU property is preserved under linear composition over an AXU has function. Thus ϕ being a linear transformation does not affect the AXU bound for aes1 and can be ignored. When L is uniformly drawn from $\mathcal{L}$ and viewed as an element of $\mathcal{D}^{16}$, then L_i's are mutually independent. So by repeatedly using Proposition 1 and Lemma 1, we get the AXU bound for aes1 in Lemma 2.

Lemma 2. *aes1 as defined above is 2^{-96}-AXU hash function.*

2) 2-Round AES based Hash Function.: From Proposition 2 we already now that the MDP for $\mathtt{aesSB}$ is upper bounded by 1.656×2^{-29}. Again we define a keyed function $\mathtt{aesSB}' : \mathbb{F}_{2^{32}} \times \mathbb{F}_{2^{32}} \to \mathbb{F}_{2^{32}}$ as

$$\forall K, x \in \mathbb{F}^2_{2^{32}}, \quad \mathtt{aesSB}'_K(x) := \mathtt{aesSB}_{K_2}(K_1 \oplus x).$$

Clearly $\mathtt{aesSB}'$ is 1.656×2^{-29}-AXU hash function (using Lemma 1). Now let $\mathcal{D} = \{0,1\}^{32}$ (viewed as $[2^{32}]$) and $\mathcal{L} = \mathcal{B} \times \mathcal{B}$ (viewed as $\mathcal{D}^4 \times \mathcal{D}^4$). Given a key $L = (L, L') \in \mathcal{D}^4 \times \mathcal{D}^4$, and an input $i \in \mathcal{D}$, we define the two round AES hash, aes2 as

$$\begin{aligned}\mathsf{aes2}_{L,L'}(i) = \mathsf{aes2}(L, L', i) &:= \mathsf{2\text{-}AESRD}_{L'}(L \oplus i^4)) \\ &= \phi(\mathtt{aesSB}'_{L'_1}(L_1 \oplus i) \| \cdots \| \mathtt{aesSB}'_{L'_4}(L_4 \oplus i)),\end{aligned}$$

where each L_i/L'_i is a 32-bit strings. Using a similar line of argument as used in case of aes1 we get the AXU bound for aes2 in Lemma 3.

Lemma 3. *aes2 as defined above is 1.881×2^{-114}-AXU hash function.*

So we see that aes1 and aes2 are sub-optimum, but decent AXU hashes given the domains they are applied upon. In fact it is quite surprising that this fact was not discovered yet. Now based on the existing proof techniques, the straightforward application of these hash functions in OCB is not advisable as they give sub-optimum security bounds.

V. Generic view of OCB

In previous two sections we have seen various examples of AXU hashes over small domains $[\ell]$, where ℓ is typically a large positive integer in $[2^n]$. These hash functions are useful when the goal is to embed position-based dependency (called masks) to the input blocks of the underlying primitive, typically a block cipher. This type of embedding is used in many encryption, MAC and AE algorithms. Some prominent schemes among these belong to the OCB like design paradigms. We call the embedding function — *mask-generating function* (MGF).

Mask Generating Function — Typically, an MGF, $\lambda : \mathcal{L} \times \mathcal{N} \times \mathbb{N}^+ \to \mathcal{B}$ is a $(\mathcal{L}, \mathcal{N} \times \mathbb{N}^+)$-hash function family indexed by the key space $\mathcal{L}$. Here $\mathcal{N}$ is typically called the nonce space. For notational simplicity we will sometime write $\lambda(i) = \lambda_L(N, i)$, i.e $N \in \mathcal{N}$ and $L \in \mathcal{L}$ will be clear from the context. Next we define generic abstraction for OCB, called GOCB, based on $\Pi \xleftarrow{\$} \mathsf{Perm}[\mathcal{B}]$ and λ. Figure 1 illustrates the schematic view of the encryption-decryption algorithms for GOCB and the complete algorithmic descriptions is given in Algorithm 1. For $i \in \mathbb{N}^+$, we will refer M_i, A_i and C_i as message block, associated data block and ciphertext block. For $i \in \mathbb{N}^+$, we will refer U_i and X_i as input blocks; V_i and Y_i as output blocks; and S^i as final block. In GOCB $X_\oplus$ and $Y_\oplus$ will be referred as the tag input and output block respectively. We will use the term GPHash[3] for the associated data processing phase of GOCB.

Note that the constructions employ different MGFs for special blocks (last block, checksum block etc.) processing. Usually for concrete constructions these MGFs are simple variants of the usual MGF λ. Similarly the key for the MGF is usually derived from the underlying random permutation. For the sake of simplicity we will assume that the MGF key is drawn independently, and the MGFs for special blocks are also defined independently.

In this paper we only focus on the mask generation component and its properties. This is mainly because GOCB greatly resemble simple variant of ECB mode when the masking is ignored. Hence both efficiency and security mostly rely on the mask-generation phase, albeit the two issues are mutually orthogonal. We discuss the two issues separately starting with efficiency.

Mask-generation and Efficiency. As noted before, if we ignore the input masking, GOCB is simply a variant of ECB, which is arguably the most efficient cryptographic mode of operation, both in hardware and software. So the MGF should preferably have the following properties:

- *Sequential Computation* — MGF should be fast enough to compute the mask outputs for consecutive inputs. More formally, for some r and for any i, $\lambda(i), \lambda(i+1), \ldots, \lambda(i+r-1)$ should be efficiently computable. Here r can be viewed as a parameter. One may observe the performance for different choices of r and then choose the best possible one.
- *Direct Computation* — MGF should be easily computable on any arbitrary input. In other words, for all i, $\lambda(i)$ should be efficiently computable without the knowledge of any other mask outputs. This is the usual constraint for hash functions. Direct computation of masks is an important feature as it leads to features like random read access and massive parallelism. Given large number of parallel processing units (multiple CPU cores or GPUs) GOCB is completely parallel when the underlying MGF is efficient in direct computation. This holds even when the message length is not known beforehand, a scenario that is common in data streams.

[3] PHash is a commonly used terminology for the hash component in OCB.

0018-9448 (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/TIT.2019.2925613, IEEE Transactions on Information Theory

Algorithm 1 GOCB authenticated encryption, where A_*, M_*, C_* are the incomplete (possibly empty) blocks at the end of A, M, C, respectively; λ_1 is a keyed (i.e. the key is implicit) mask-generating function, and λ_i for $1 < i \leq 6$ are minor variants of λ_1, deployed in various boundary conditions, like incomplete block processing, tag generation etc.

function $\mathsf{GOCB}[\Pi, \lambda].\mathsf{Enc}(N, A, M)$
- $S \leftarrow 0^n$
- $M_{\oplus} \leftarrow 0^n$
- $(A_1, \ldots, A_m, A_*) \xleftarrow{n} A$
- $(M_1, \ldots, M_\ell, M_*) \xleftarrow{n} M$
- **for** $i \leftarrow 1$ **to** m **do**
  - $U_i \leftarrow A_i \oplus \lambda_5(N, i)$
  - $V_i \leftarrow \Pi(U_i)$
  - $S \leftarrow S \oplus V_i$
- **if** $|A_*| > 0$ **then**
  - $U_* \leftarrow \mathsf{pad}(A_*) \oplus \lambda_6(N, m)$
  - $V_* \leftarrow \Pi(U_*)$
  - $S \leftarrow S \oplus V_*$
- **for** $i \leftarrow 1$ **to** ℓ **do**
  - $M_{\oplus} \leftarrow M_{\oplus} \oplus M_i$
  - $X_i \leftarrow M_i \oplus \lambda_1(N, \ell)$
  - $Y_i \leftarrow \Pi(X_i)$
  - $C_i \leftarrow Y_i \oplus \lambda_1(N, i)$
- **if** $|M_*| > 0$ **then**
  - $M_{\oplus} \leftarrow M_{\oplus} \oplus \mathsf{pad}(M_*)$
  - $X_* \leftarrow \lambda_2(N, \ell)$
  - $Y_* \leftarrow \Pi(X_*)$
  - $\overline{C_*} \leftarrow Y_* \oplus \mathsf{pad}(M_*)$
  - $C_* \leftarrow \mathsf{msb}_{|M_*|}(\overline{C_*})$
  - $C \leftarrow (C_1, \ldots, C_\ell, C_*)$
  - $X_{\oplus} \leftarrow M_{\oplus} \oplus \lambda_4(N, \ell)$
  - $Y_{\oplus} \leftarrow \Pi(X_{\oplus})$
  - $T \leftarrow Y_{\oplus} \oplus \lambda_4(N, \ell) \oplus S$
- **else**
  - $C \leftarrow (C_1, \ldots, C_\ell)$
  - $X_{\oplus} \leftarrow M_{\oplus} \oplus \lambda_3(N, \ell)$
  - $Y_{\oplus} \leftarrow \Pi(X_{\oplus})$
  - $T \leftarrow Y_{\oplus} \oplus \lambda_3(N, \ell) \oplus S$
- **return** (C, T)

function $\mathsf{GOCB}[\Pi, \lambda].\mathsf{Dec}(N, A, C, T)$
- $S \leftarrow 0^n$
- $M_{\oplus} \leftarrow 0^n$
- $(A_1, \ldots, A_m, A_*) \xleftarrow{n} A$
- $(C_1, \ldots, C_\ell, C_*) \xleftarrow{n} C$
- **for** $i \leftarrow 1$ **to** m **do**
  - $U_i \leftarrow A_i \oplus \lambda_5(N, i)$
  - $V_i \leftarrow \Pi(U_i)$
  - $S \leftarrow S \oplus V_i$
- **if** $|A_*| > 0$ **then**
  - $U_* \leftarrow \mathsf{pad}(A_*) \oplus \lambda_6(N, m)$
  - $V_* \leftarrow \Pi(U_*)$
  - $S \leftarrow S \oplus V_*$
- **for** $i \leftarrow 1$ **to** ℓ **do**
  - $Y_i \leftarrow C_i \oplus \lambda_1(N, \ell)$
  - $X_i \leftarrow \Pi^{-1}(Y_i)$
  - $M_i \leftarrow X_i \oplus \lambda_1(N, i)$
  - $M_{\oplus} \leftarrow M_{\oplus} \oplus M_i$
- **if** $|C_*| > 0$ **then**
  - $X_* \leftarrow \lambda_2(N, \ell)$
  - $Y_* \leftarrow \Pi(X_*)$
  - $M_* \leftarrow \mathsf{msb}_{|C_*|}(Y_*) \oplus C_*$
  - $M \leftarrow (M_1, \ldots, M_\ell, M_*)$
  - $M_{\oplus} \leftarrow M_{\oplus} \oplus \mathsf{pad}(M_*)$
  - $X_{\oplus} \leftarrow M_{\oplus} \oplus \lambda_4(N, \ell)$
  - $Y_{\oplus} \leftarrow \Pi(X_{\oplus})$
  - $T' \leftarrow Y_{\oplus} \oplus S$
- **else**
  - $M \leftarrow (M_1, \ldots, M_\ell)$
  - $X_{\oplus} \leftarrow M_{\oplus} \oplus \lambda_3(N, \ell)$
  - $Y_{\oplus} \leftarrow \Pi(X_{\oplus})$
  - $T' \leftarrow Y_{\oplus} \oplus S$
- **if** $T' = T$ **then**
  - **return** M
- **else**
  - **return** $\perp$

- *Constant-time Computation* — A secondary requirement for any MGF is constant computation time for all inputs. While this is a secondary requirement as far as efficiency is concerned, it helps in mitigating a form of side channel attacks called timing attack [51], [33].

MASK-GENERATION AND SECURITY — From the security point of view, earlier works [3], [9], [37] required strong (almost 2^{-n}) AXU and 1-universal (regular, i.e $\Pr[H(x) = y] = O(2^{-n})$) property (some time implicitly) from the underlying MGFs. Since allmost all of the hash functions discussed in this work are regular, we will implicitly assume that MGF has this property. We will focus on the more dominant, AXU property. In this setting the advantage is generally bounded in terms of ϵ, the AXU bound of MGF. For instance the bound for OCB is roughly $O(\sigma^2\epsilon)$ respectively; hence the need for $O(2^{-n})$-AXU bound. Naturally, both aes1 and aes2 would suffer significant security loss in the existing setting. This is probably one of the reasons, these hash functions have not seen much attention. Yet another reason is the fact that aes1 and aes2 are defined over restricted domains $\{0,1\}^8$ and $\{0,1\}^{32}$ respectively.

A. Locally-Imperfect XOR Universal Hash Functions

The immediate question in light of the above discussion is as follows:

> *Can we relax the AXU condition on MGFs to use efficient directly computable functions like* aes1 *and* aes2*, maintaining comparable security?*

We answer this in affirmative by proposing a slightly different universal notion — *Locally-Imperfect XOR Universal (or LIXU) Hash Functions*. The idea is to model a hash function H whose differential probability depends on the grouping of the given inputs. More precisely, we partition the inputs in several disjoint subsets with bounded cardinality, say at most r.[4] Now the additional property introduced by LIXU is as follows: for two inputs say x and y from distinct partitions the differential probability is at most 2^{-n}, and the probability degrades to $\epsilon \geq 2^{-n}$, when x and y belong to the same partition. Formally we define the notion of Locally-Imperfect XOR Universal in Definition 3.

Definition 3 (Locally-Imperfect XOR Universal Hash Function)**.** Let $H : \mathcal{L} \times \mathcal{D} \to \mathcal{B}$ be a $(\mathcal{L}, \mathcal{D})$-hash function family indexed by the keyspace $\mathcal{L}$. For a fixed $r \in \mathbb{N}^+$ and $\epsilon \geq 2^{-n}$,

[4] In fact we try to minimize the value of r.

0018-9448 (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/TIT.2019.2925613, IEEE Transactions on Information Theory

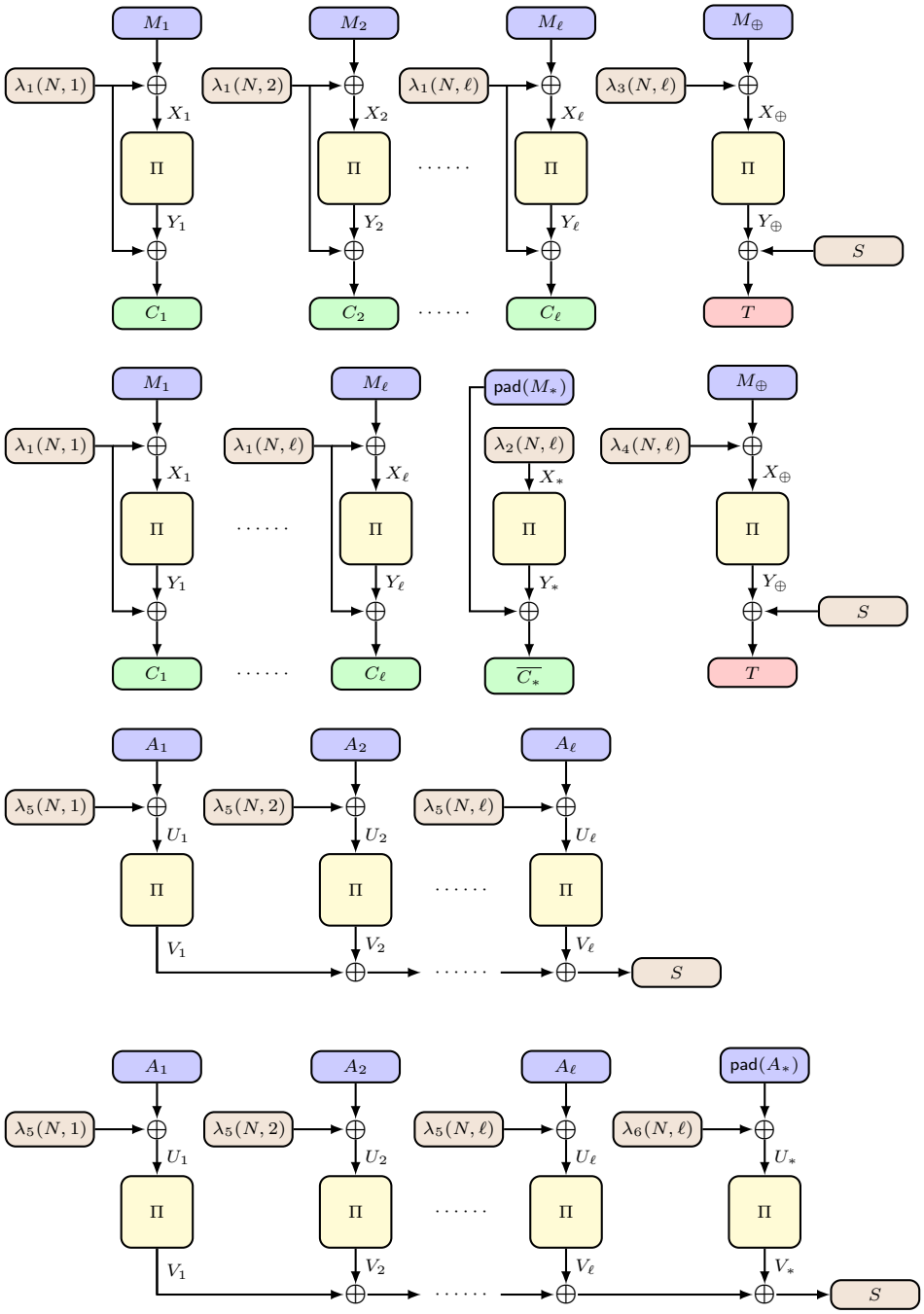


Fig. 1: Schematic diagram of GOCB— **Top to Bottom:** encrypting M when $|M_*| = 0$; encrypting M when $|M_*| > 0$; hashing A when $|A_*| = 0$; hashing A when $|A_*| > 0$.

we say that H is an (ϵ, r)-locally-imperfect XOR universal, or LIXU, hash function if and only if there exists a partitioning $\mathcal{D} = \mathcal{P}_1 \sqcup \mathcal{P}_2 \sqcup \cdots \sqcup \mathcal{P}_k$ with $k \geq 2$, such that $|\mathcal{P}_i| \leq r$ for $i \in [k]$, and for distinct $x, y \in \mathcal{D}$, and $\delta \in \mathcal{B}$,

$$\Pr_{L \xleftarrow{\$} \mathcal{L}}[H_L(x) + H_L(y) = \delta] \leq \begin{cases} \epsilon & \text{if } x, y \in \mathcal{P}_i, \\ \frac{1}{2^n} & \text{if } x \in \mathcal{P}_i,\ y \in \mathcal{P}_j, i \neq j. \end{cases}$$

Here r is called the *width* of H. We say that $x, y \in \mathcal{D}$ are *local pairs* (to each other and as a pair as well) if and only if they belong to the same partition. So a LIXU hash function behaves as a perfect XOR universal hash function when the inputs are not local pairs, and behaves as an imperfect (or almost) XOR universal hash function when the inputs are local pairs. Note that an (ϵ, r)-LIXU hash implies an ϵ-AXU hash, but the vice-versa may not be true. Further it is rather trivial to establish that any 2^{-n}-AXU hash function over the domain $\mathcal{D}$ is also a $(2^{-n}, r)$-LIXU hash function for all $r \leq |\mathcal{D}| - 1$. We note this simple fact in Proposition 3.

Proposition 3. *A 2^{-n}-AXU hash H is a $(2^{-n}, r)$-LIXU hash for all $r \leq |\mathcal{D}| - 1$, where $\mathcal{D}$ is the domain of H. Notably, H*

0018-9448 (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/TIT.2019.2925613, IEEE Transactions on Information Theory

is a $(2^{-n}, 1)$-LIXU hash function.

In this paper, we use a specific type of LIXU hash functions defined over the domain $\mathcal{N} \times \mathbb{N}^+$, where the domain is partitioned as follows: $(N, i) \neq (N', i') \in \mathcal{N} \times \mathbb{N}^+$ are said to be local pairs if and only if $N = N'$ and $\lceil i/r \rceil = \lceil i'/r \rceil$. Here, $\mathcal{N}$ denotes the nonce space. We sometime use chunk and partition interchangeably, and for any $i \in \mathbb{N}^+$, we call $\lceil i/r \rceil$ the chunk number of i.

B. LIXU Hash as MGF in GOCB

Previous works on OCB have assumed that the underlying MGF is an AXU hash. We depart from this setting and employ LIXU hash as MGF. Since LIXU and AXU are not exactly compatible, we will require a slightly different analysis. Particularly the TBC based abstraction [8] for OCB is not useful here. We now present the main security result of the paper, along with an overview of our proof approach. The formal proof is postponed to Section VI.

Theorem 1 (GOCB NAEAD Bound)**.** *For a fixed $r \leq \ell \in \mathbb{N}^+$ and $\epsilon \geq 0$, let λ be an (ϵ, r)-LIXU hash function over $\mathcal{N} \times \mathbb{N}^+$. Let $\mathbb{A}(q, \tilde{q}, \ell, \sigma, \tilde{\sigma})$ be the class of all adversaries that make q encryption queries consisting of μ message blocks in all with at most ℓ blocks per query, and ν associated data blocks in all, and $\tilde{q}$ decryption queries in a NAEAD security game against $\mathsf{GOCB}[\Pi, \lambda]$. Then $\forall \mathscr{A} \in \mathbb{A}(q, \tilde{q}, \ell, \sigma)$ we have, $\mathbf{Adv}^{\mathsf{naead}}_{\mathsf{GOCB}[\Pi,\lambda]}(\mathscr{A}) = \epsilon_{\mathsf{naead}}$, where*

$$\epsilon_{\mathsf{naead}} \leq \frac{\sigma^2}{2^{n+1}} + \frac{q\sigma + \tilde{q} + 0.5q^2}{2^n - \sigma} + \frac{\tilde{\sigma}(\sigma+4)}{2^n} + (\sigma + 3\tilde{\sigma})r\epsilon + \tilde{q}\epsilon.$$

Here $\sigma = \mu + \nu$ is the total number of blocks in messages and associated data queried in the encryption queries.

Proof Sketch — The proof of Theorem 1 is given in Section VI. While we do get a factor of $r\sigma\epsilon$ in the bound for GOCB, the original bound of OCB remains intact asymptotically if $r\epsilon = o(2^{\frac{n}{2}})$. To understand why we don't get any degradation in security we have to look into the bound more precisely. Generally the OCB bound is dominated by the probability of collision among the input/output of the underlying block cipher. As per the earlier analyses the collisions can be bounded in terms of the AXU bound of λ, which gives $\sigma^2\epsilon/2$ bound. This evaluates to $\sigma^2 2^{1-n}$ when $\epsilon = 2^{-n}$. However when we replace AXU with LIXU the straightforward analysis will result in security degradation. Instead we first bound the probability of collision between input blocks which are non-local pairs. Since LIXU hashes are perfectly XOR universal on non-local pairs we get a bound of $\sigma^2 2^{1-n}$. Now for each input block there are at most $r-1$ local pairs and there are at most σ many such input blocks which gives a bound of roughly $r\sigma\epsilon$. Following similar idea we can bound the probability of forgery to $\tilde{q}\sigma 2^{-n} + \tilde{q}r\epsilon$.

C. Instantiating LIXU hash in GOCB

1) Deriving OCB3 from GOCB: We define the mask-generating function used in OCB3 in LIXU setting and derive concrete bound for $n = 128$. OCB3 employs gray code based masking defined using the gray hash function discussed in Section III. As noted earlier gray is a 2^{-128}-AXU hash function.

For OCB3, Krovetz and Rogaway defined the gray function as: For any $i \in \mathbb{N}^+$, $\mathsf{gray}_L(i) = 4\gamma(i) \odot L$. For a fixed $j \in [n]$ (fixed as an application parameter) and $N \in \mathcal{B} \setminus \{0\}$, $\tau(N) = \mathsf{msb}_j(N) \| 0^{n-j}$, and $\beta(N) = \mathsf{lsb}_j(N)$. The OCB3 NAEAD advantage [9] can be obtained from GOCB by instantiating the definition of GOCB as follows:

- $L = (L_1, L_2)$, where L_2 is a keyed function from $\mathcal{B}$ to $\mathcal{B}$, with $L_1 := E_K(0)$ and $\forall N \in \mathcal{B}$, $L_2(N) := \mathsf{sts}_{E_K(\tau(N))}(\beta(N))$.
- For $i \in [2^n]$, and $N \in \mathcal{B} \setminus \{0\}$:
  - $\lambda_1(L, N, i) := L_2(N) \oplus \mathsf{gray}_{L_1}(i)$; $\lambda_5(L, N, i) := \mathsf{gray}_{L_1}(i)$;
  - $\lambda_2(L, N, i) := L_2(N) \oplus \mathsf{gray}_{L_1}(i) \oplus L_1$; $\lambda_6(L, N, i) := \mathsf{gray}_{L_1}(i) \oplus L_1$;
  - $\lambda_3(L, N, i) := L_2(N) \oplus \mathsf{gray}_{L_1}(i) \oplus 2L_1$.
  - $\lambda_4(L, N, i) := L_2(N) \oplus \mathsf{gray}_{L_1}(i) \oplus 3L_1$.

Here we have derived L through K, which adds a insignificant factor in the overall advantage. In [9], all variants of GR were shown to be 2^{-n}-AXU hash over the domain $\{0,1\}^{128}$. This in combination with Proposition 3, establishes that GR variants are $(2^{-n}, 1)$-LIXU hash. Using Theorem 1 and assuming $\sigma, \widetilde{\sigma} < 2^{n-1}$, we get the security bound for the original OCB3 design [9].

Corollary 1. *$\forall \mathscr{A} \in \mathbb{A}(q, \ell, \sigma)$ we have,*

$$\mathbf{Adv}^{\mathsf{naead}}_{\mathsf{OCB3}[E_K^{\pm},\mathsf{GR}]}(\mathscr{A}) \leq \mathbf{Adv}^{\mathsf{sprp}}_{E_K^{\pm}}(\sigma + \widetilde{\sigma} + q + \widetilde{q}) + \epsilon_{\mathsf{GR}},$$

where

$$\epsilon_{\mathsf{GR}} = \frac{0.5\sigma^2 + 2q\sigma + q^2 + \widetilde{\sigma}\sigma + \sigma + 7\widetilde{\sigma} + 3\widetilde{q}}{2^n}.$$

2) Instantiating GOCB with aes1 and aes2: We set $n = 128$. We let $\mathcal{N} \subset \mathcal{B}$, and define

- $\mathsf{A1} : \mathsf{Func}[\mathcal{B}] \times \mathcal{N} \times \mathbb{F}_{2^8} \to \mathcal{B}$ as, $\forall F \in \mathsf{Func}[\mathcal{B}]$, $N \in \mathcal{N}$, $i \in \mathbb{F}_{2^8}$,
$$\mathsf{A1}_F(N, i) := \mathsf{aes1}_{F(N)}(i).$$
- $\mathsf{A2} : \mathsf{Func}[\mathcal{B}] \times \mathcal{B} \times \mathcal{N} \times \mathbb{F}_{2^{32}} \to \mathcal{B}$ as, $\forall (F, L, N, i) \in \mathsf{Func}[\mathcal{B}] \times \mathcal{B} \times \mathcal{N} \times \mathbb{F}_{2^{32}}$,
$$\mathsf{A2}_{(F,L)}(N, i) := \mathsf{aes2}_{(F(N),L)}(i).$$

Next we show that A1 and A2 are LIXU hash functions for appropriate values of ϵ and r. In Lemma 4 and Lemma 5 we show that A1 and A2 are $(2^{-96}, 2^8)$-LIXU and $(2^{-113}, 2^{32})$-LIXU hash functions, respectively. The proofs of these lemmata are given in Section VI.

Lemma 4. *If $\Gamma \leftarrow_{\$} \mathsf{Func}[\mathcal{B}]$, then $\mathsf{A1}_\Gamma$ is a $(2^{-96}, 2^8)$-LIXU hash function.*

Lemma 5. *If $(\Gamma \times L) \leftarrow_{\$} \mathsf{Func}[\mathcal{B}] \times \mathcal{B}$, then $\mathsf{A2}_{\Gamma,L}$ is a $(2^{-113}, 2^{32})$-LIXU hash function.*

0018-9448 (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/TIT.2019.2925613, IEEE Transactions on Information Theory

DESCRIPTION OF $\mathsf{GOCB}[E, \mathsf{A1}]$ — We define GOCB based on $\mathsf{A1}$ as follows:

- The maximum message and associated data size per query is limited to 2^6 blocks, i.e. 1 Kilobytes, and the nonce space is $\mathcal{B} \setminus \{0\}$.
- $L = F$, where F is a keyed function from $\mathcal{B}$ to $\mathcal{B}$, where $\forall N \in \mathcal{B}$, $F(N) := E_K(N)$.
- For $i \in [2^6]$, and $N \in \mathcal{B} \setminus \{0\}$:
  - $\lambda_1(F, N, i) := \mathsf{A1}_F(N, 4i + 0)$; and $\lambda_5(F, N, i) := \mathsf{A1}_F(N, 4i + 0) \oplus E_K(0)$.
  - $\lambda_2(F, N, i) := \mathsf{A1}_F(N, 4i + 1)$; and $\lambda_6(F, N, i) := \mathsf{A1}_F(N, 4i + 1) \oplus E_K(0)$.
  - $\lambda_3(F, N, i) := \mathsf{A1}_F(N, 4i + 2)$;
  - $\lambda_4(F, N, i) := \mathsf{A1}_F(N, 4i + 3)$;

  where $4i$ denotes the integer multiplication of i with 4.

Here we need to add a factor of $(q+1)(q+\sigma)2^{-128}$ to bound the probability that some masked input matches with some nonce value. Further a factor of $q^2 2^{-129}$ is required for PRF-PRP switch. Using Theorem 1, and $\sigma, \widetilde{\sigma} < 2^{127}$, we get the security result for $\mathsf{GOCB}[E_K, \mathsf{A1}]$ as in Corollary 2.

Corollary 2. *$\forall \mathscr{A} \in \mathbb{A}(q, \tilde{q}, \sigma)$ we have,*

$$\mathbf{Adv}^{\mathsf{naead}}_{\mathsf{GOCB}[E_K^{\pm}, \mathsf{A1}]}(\mathscr{A}) \leq \mathbf{Adv}^{\mathsf{sprp}}_{E_K^{\pm}}(\sigma + q) + \epsilon_{\mathsf{A1}},$$

where

$$\epsilon_{\mathsf{A1}} = \frac{0.5\sigma^2 + 6q\sigma + 2\sigma + \widetilde{\sigma}\sigma + 4\widetilde{\sigma} + 2\tilde{q}}{2^{128}} + \frac{\sigma + 3\widetilde{\sigma}}{2^{88}} + \frac{\widetilde{q}}{2^{96}}.$$

DESCRIPTION OF $\mathsf{GOCB}[E, \mathsf{A2}]$ — $\mathsf{GOCB}[E, \mathsf{A2}]$ can be defined analogously using mask hash key $L = (L_1, L_2, F)$, where $L_1 = E_K(0)$ and $L_2 = E_K(1)$. We define GOCB based on $\mathsf{A2}$ as follows:

- The maximum message and associated data size per query is limited to 2^{30} blocks, i.e. 16 Gigabytes, and the nonce space is $\mathcal{B} \setminus \{0, 1\}$.
- $L = (L_1, F)$, where $L_1 = E_K(0)$ and F is a keyed function from $\mathcal{B}$ to $\mathcal{B}$. For all $N \in \mathcal{B} \setminus \{0, 1\}$, $F(N) := E_K(N)$.
- For $i \in [2^{30}]$, and $N \in \mathcal{B}$:
  - $\lambda_1(L, N, i) := \mathsf{A2}_L(N, 4i + 0)$; and $\lambda_1(L, N, i) := \mathsf{A2}_L(N, 4i + 0) \oplus E_K(1)$.
  - $\lambda_2(L, N, i) := \mathsf{A2}_L(N, 4i + 1)$; and $\lambda_2(L, N, i) := \mathsf{A2}_L(N, 4i + 1) \oplus E_K(1)$.
  - $\lambda_3(L, N, i) := \mathsf{A2}_L(N, 4i + 2)$;
  - $\lambda_4(L, N, i) := \mathsf{A2}_L(N, 4i + 3)$;

  where $4i$ denotes the integer multiplication of i with 4.

Using similar argument as above we can derive the security bound, given in Corollary 3.

Corollary 3. *$\forall \mathscr{A} \in \mathbb{A}(q, \tilde{q}, \sigma)$ we have,*

$$\mathbf{Adv}^{\mathsf{naead}}_{\mathsf{GOCB}[E_K^{\pm}, \mathsf{A2}]}(\mathscr{A}) \leq \mathbf{Adv}^{\mathsf{sprp}}_{E_K^{\pm}}(\sigma + q) + \epsilon_{\mathsf{A2}},$$

where

$$\epsilon_{\mathsf{A2}} = \frac{0.5\sigma^2 + 6q\sigma + 4\sigma + \widetilde{\sigma}\sigma + 4\widetilde{\sigma} + 2\tilde{q}}{2^{128}} + \frac{\sigma + 3\widetilde{\sigma}}{2^{81}} + \frac{\widetilde{q}}{2^{113}}.$$

3) Extending the Domain of GOCB[E,A1] *and* GOCB[E,A2]*:* The $\mathsf{A1}$ and $\mathsf{A2}$ based instantiations have a restriction on the maximum input length. This is more prominent in case of $\mathsf{GOCB}[E, \mathsf{A1}]$, which allows message lengths up to 1 KB. Here we give some ways to extend the domain of $\mathsf{GOCB}[E, \mathsf{A1}]$ and $\mathsf{GOCB}[E, \mathsf{A2}]$. Our general idea remains the same in all the methods. We improve the maximum input size restriction to roughly 2^{50} bytes or 2^{46} blocks, a usual limit on input size [52]. We need 46 bits to represent each block of input uniquely. We divide the 46-bit length representation into two parts $i \| j$, where j is 6-bit long for $\mathsf{GOCB}[E, \mathsf{A1}]$ and j is 30-bit long for $\mathsf{GOCB}[E, \mathsf{A2}]$. We will handle the j-value in the same way as before, and employ different techniques to handle (N, i) input of the mask generating function.

1. DOMAIN EXTENSION VIA THE F FUNCTION: We can consider a keyed function over larger domain, say $\mathcal{N} \times [2^{128}]$. For example, consider the definition $F(N, i) := E_K(i \oplus E_K(N))$. This would add an extra $O(\sigma^2/2^n)$ term, a rough upper bound to avoid block cipher input collisions involving $i \oplus E_K(N)$ for some (N, i). This method can clearly extend the message size to 2^{46} but at the cost of an extra block cipher call every time the j value resets to zero, plus it needs one more encryption call for the nonce.

2. DOMAIN EXTENSION VIA NONCE SIZE REDUCTION: Another way is to reduce the nonce size to 88 bits and 112 bits for $\mathsf{GOCB}[E, \mathsf{A1}]$ and $\mathsf{GOCB}[E, \mathsf{A2}]$, respectively, and let $F(N, i) := E_K(N \| i)$. This is slightly better than the previous method as we can save one block cipher call. As before, this adds an extra $O(\sigma^2/2^n)$ term, a rough upper bound to avoid block cipher input collisions.

3. DOMAIN EXTENSION VIA SUM OF LIXU HASH: We explain the method for $\mathsf{GOCB}[E, \mathsf{A2}]$. But similar technique is also applicable to $\mathsf{GOCB}[E, \mathsf{A1}]$. We modify the λ function as follows:

- First, we redefine $L = (L_1, L_2, F, F')$, where $L_1 = E_K(0)$ and $L_2 = E_K(1)$. The nonce space is $\mathcal{N} = \{0, 1\}^{n-1} \setminus \{0, 1\}$, and F and F' are keyed functions from $\mathcal{N}$ to $\mathcal{B}$, defined as $F_K(N) = E_K(N \| 0)$ and $F'_K(N) = E_K(N \| 1)$.
- Second, $\lambda_1(L, N, i \| j) := \mathsf{A2}_{L_1, F}(N, i) \oplus \mathsf{A2}_{L_2, F}(N, j)$. The other variants are defined analogously as before. Further for AD processing, we XOR $E_K(2)$ (instead of $E_K(1)$) to the modified λ_1 and λ_2 values.

It can be easily shown that this modification offers asymptotically similar security as before.

0018-9448 (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/TIT.2019.2925613, IEEE Transactions on Information Theory

VI. Security Proofs

In this section we provide the proof for our main result Theorem 1, along with the proofs for Lemma 4 and Lemma 5.

A. Inaptness of the Existing TBC-based Abstraction

In [8], Rogaway introduced a very nice abstraction for OCB like constructions based on tweakable block ciphers (TBCs). More formally, the TBC abstraction of [8] views the masked input-output block cipher as a way to construct TBCs from block ciphers and efficient mask generating functions, called XEX. In this view any OCB like construction can be viewed as an instance of the TBC-based authenticated encryption called ΘCB, where each encryption TBC call takes a tuple of tweak values (N, i), where N denotes the nonce and i denotes the block index.

This abstraction simplifies as well as modularizes the proof, as it can be shown (using results from [8]) that the NAEAD advantage is bounded by the tweakable strong pseudorandom permutation (TSPRP) advantage of XEX, which is bounded by at most $\sigma^2\epsilon$, where σ denotes the total number of TBC calls and ϵ denotes the AXU bound of the underlying MGF. Consequently, this approach was used in several previous works, most notably [27] and [26].

While the reduction from the NAEAD game for OCB to the TSPRP game for XEX gives tight estimate on the privacy bounds (exactly same as the TSPRP bound for XEX), the reduction is rather loose in case of authenticity bound, as noted in [37]. A straightforward approach gives an authenticity bound of the form $O(\sigma^2\epsilon)+O(1/2^n)$ for single forgery, where σ denotes the number of TBC calls in all encryption queries. The first term here is due to the TSPRP advantage of XEX. Here the $O(\sigma^2\epsilon)$ term is due to the XEX to tweakable random permutation reduction. When extended to multiple verification queries, this results in a bound of the form $O(\tilde{q}\sigma^2\epsilon)+O(\tilde{q}/2^n)$, where $\tilde{q}$ denotes the number of verification queries. Clearly the security degrades for multiple verification queries.

Yet another approach gives a security bound of the form $O(\sigma^2/\epsilon) + O(\tilde{\sigma}^2\epsilon) + O(\tilde{q}/2^n)$, where $\tilde{\sigma}$ denotes the number of TBC calls in all decryption queries. Here the $O(\sigma^2/\epsilon)$ and $O(\tilde{\sigma}^2\epsilon)$ terms are due to the XEX to tweakable random permutation reduction.

A constrained variant of the XEX based abstraction could also be used to get a modularized privacy bound for GOCB. In this abstraction the adversary against tweakable block cipher is restricted to at most r many queries per chunk number. This restriction is perfectly fine, as the NAEAD adversary is nonce-respecting for encryption queries and for each nonce the message length is at most r. This gives a privacy bound of the form $O(\sigma^2/2^n) + O(\sigma r\epsilon)$. But the same is not true for authenticity, as the adversary can make all the decryption queries with tweaks within the same chunk. This results in a term of the form $O(\tilde{\sigma}^2\epsilon)$, which is clearly sub-optimal.

So, we use a more direct approach as also employed in [37]. The proof for Theorem 1 uses coefficient H technique [53] as the main tool for bounding the distinguishing advantages. So, before delving into the proofs we discuss this tool briefly.

B. Coefficient H Technique

We will consider a computationally unbounded and deterministic adversary $\mathscr{A}$ that tries to distinguish $\mathcal{O}_1$ (the real oracle) from $\mathcal{O}_0$ (the ideal oracle) in a security game, in this case NAEAD. We denote the query-response tuple of $\mathscr{A}$'s interaction with its oracle by a transcript ω. Sometime this may also include any additional information the oracle chooses to reveal to the adversary at the end of the query-response phase of the game. We will consider this extended definition of transcript. The probability of realizing a given transcript ω in the security game with an oracle $\mathcal{O}$ is known as the *interpolation probability* of ω with respect to $\mathcal{O}$, denoted $\mathsf{ip}_{\mathcal{O}}[\omega]$. Note that for a transcript to be realized, two things need to happen:

- The adversary needs to make the queries listed in the transcript;
- The oracle needs to make the corresponding responses.

Of these, the former is deterministic; the latter, probabilistic. Thus when we talk of interpolation probability, we are only concerned with the oracle responses, with the assumption that the adversary's queries are consistent with the transcript. For any other adversary, the interpolation probability is trivially 0. Thus $\mathsf{ip}_{\mathcal{O}}[\omega]$ depends only on the oracle $\mathcal{O}$ and the transcript ω and not on the adversary; hence the notation. A transcript ω is said to be *realizable* if $\mathsf{ip}_{\mathcal{O}_0}[\omega] > 0$.
We extend the notation of interpolation probability to a set of transcripts Ω: $\mathsf{ip}_{\mathcal{O}}[\Omega]$ is the probability that the security game with $\mathcal{O}$ results in a transcript $\omega \in \Omega$, and we say that $\mathsf{ip}_{\mathcal{O}}[\Omega]$ is the interpolation probability of Ω with respect to $\mathcal{O}$. Now we state a theorem due to Patarin, known as the Coefficient H Technique [53].

Theorem 2 (Coefficient H Technique [53], [54])**.** *Let Ω be the set of all realizable transcripts. Suppose there is a set $\Omega_{\mathtt{Bad}} \subseteq \Omega$ satisfying the following:*

- $\mathsf{ip}_{\mathcal{O}_0}[\Omega_{\mathtt{Bad}}] \leq \epsilon_{\mathsf{bad}}$;
- *For any $\omega \notin \Omega_{\mathtt{Bad}}$,*

$$\frac{\mathsf{ip}_{\mathcal{O}_1}[\omega]}{\mathsf{ip}_{\mathcal{O}_0}[\omega]} \geq 1 - \epsilon_{\mathsf{ratio}}.$$

Then for an adversary $\mathscr{A}$ trying to distinguish between $\mathcal{O}_1$ and $\mathcal{O}_0$, we have the following bound on its distinguishing advantage:

$$\mathbf{Adv}^{G}_{\mathcal{O}_1}(\mathscr{A}) \leq \epsilon_{\mathsf{bad}} + \epsilon_{\mathsf{ratio}}.$$

Note that we have dropped $\mathcal{O}_0$ from the subscript, as it is completely determined by G. A proof of this theorem is given in multiple papers including [54], [55], [56]. A weaker version of the result was later rediscovered by Bernstein [57] as the *interpolation theorem*, which was later strengthened by Nandi [58] as the *strong interpolation theorem*.

Additional Notation — For a vector x, $\#x_i$ denotes the multiplicity of x_i in x (number of $j \in |x|$ such that $x_j = x_i$). We sometime abuse the notation and view x as the set $\{x_i : i \in [|x|]\}$, and denote the cardinality of this set as $\#x$. For $i \in \mathbb{N}^+$, if x is a tuple of i vectors, then x^i denotes the

0018-9448 (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/TIT.2019.2925613, IEEE Transactions on Information Theory

i-th vector in the tuple and for some $j \in [|x^i|]$, x^i_j denotes the j-th coordinate of x^i.

We say a function $f : \mathcal{D} \to \mathcal{R}$ is *partial* or *restricted*, if we know the values of f on a strict subset of $\mathcal{D}$. This subset is called $\mathsf{domain}(f)$. A partial function f can be viewed as a restriction of f to $\mathsf{domain}(f)$. The range of this restricted function is called $\mathsf{range}(f)$. We will treat a partial function as *updatable*: for some $x \in \mathcal{B} \setminus \mathsf{domain}(f)$ and some $y \in \mathcal{R}$, (x, y) may be added to f, so that $\mathsf{domain}(f)$ expands to $\mathsf{domain}(f) \cup \{x\}$, and $\mathsf{range}(f)$ becomes $\mathsf{range}(f) \cup \{y\}$. We say f is *permutation-compatible* if $|\mathsf{domain}(f)| = |\mathsf{range}(f)|$ and f is injective function on the set it is defined.

C. Proof of GOCB NAEAD Bound (Theorem 1)

Let $\mathcal{O}_1^{\pm}$ denote the real oracle corresponding to GOCB, and $\mathcal{O}_0^{\pm}$ denote the ideal oracle corresponding to $(\Gamma, \perp)$. We start off by setting up the necessary notations related to the query-response transcript of $\mathscr{A}$. The transcript generated by $\mathscr{A}$ consists of

- **Encryption query-response tuples:** for $i \in [q]$, the i-th encryption query-response tuple is built of
  - N^i: the queried nonce block, such that for any $i' < i$, $N^i \neq N^{i'}$.
  - A^i: the queried associated data, consisting of a^i complete blocks and an incomplete (possibly empty) block A^i_* at the end;
  - M^i: the queried message, consisting of m^i complete blocks and an incomplete (possibly empty) block M^i_* at the end;
  - C^i: the response ciphertext, such that $|C^i| = |M^i|$;
  - T^i: the response tag block.
- **Decryption query-response tuples:** for $i \in [\widetilde{q}]$ the i-th decryption query-response tuple is built of
  - $\widetilde{N}^i$: the queried nonce block. Note that $\widetilde{N}^i$'s can be repeated.
  - $\widetilde{A}^i$: the queried associated data, consisting of $\widetilde{a}^i$ complete blocks and an incomplete (possibly empty) block $\widetilde{A}^i_*$ at the end;
  - $\widetilde{C}^i$: the queried ciphertext, consisting of $\widetilde{c}^i$ complete blocks and an incomplete (possibly empty) block $\widetilde{C}^i_*$ at the end;
  - $\widetilde{T}^i$: the queried tag block.
  - $\widetilde{M}^i$: the response message, such that $|\widetilde{M}^i| = |\widetilde{C}^i|$. Note that the decryption oracle may return $\perp$, in which case $\widetilde{M}^i$ just denotes the unauthenticated decrypted message to be used internally.

The internal variables arising in one call to the encryption oracle are analogously as given in Algorithm 1 and Figure 1, while the internal variables from the decryption oracle are defined identically, but topped with a tilde to differentiate them from the encryption variables.

Let $\mathcal{I}$ and $\mathcal{J}$ denote the encryption query indices with incomplete-block message and associated data respectively. The decryption counterparts are $\widetilde{\mathcal{I}}$ and $\widetilde{\mathcal{J}}$. Let I and O be multisets defined as

$$I = \{U^i_j \ : \ i \in [q], \ j \in [a^i]\} \cup \{U^i_* \ : \ i \in \mathcal{J}\} \cup \{X^i_j \ : \ i \in [q], \ j \in [m^i] \cup \{\oplus\}\} \cup \{X^i_* \ : \ i \in \mathcal{I}\}$$

$$O = \{V^i_j \ : \ i \in [q], \ j \in [a^i]\} \cup \{V^i_* \ : \ i \in \mathcal{J}\} \cup \{Y^i_j \ : \ i \in [q], \ j \in [m^i] \cup \{\oplus\}\} \cup \{Y^i_* \ : \ i \in \mathcal{I}\}$$

1) Oracle behavior and transcript extension: Consider a modified security game where we let the real oracle $\mathcal{O}_1$ reveal all the internal input and output blocks appearing in the encryption query phase. Consequently we have to make appropriate behavioral changes in the ideal oracle $\mathcal{O}_0$ to release these internal variables. We describe the sampling behavior of $\mathcal{O}_0$ in greater detail in the subsequent paragraphs. During the sampling $\mathcal{O}_0$ might set **bad** $= 1$, whereafter its behavior is undefined.

SAMPLING BEHAVIOR OF $\mathcal{O}_0$

Query Phase: For $i \in [q]$, on the i-th encryption query, for each $j \in [m^i]$, $C^i_j \leftarrow\!\!\$\ \mathcal{B}$; $T^i \leftarrow\!\!\$\ \mathcal{B}$; and if $i \in \mathcal{I}$, $\overline{C^i_*} \leftarrow\!\!\$\ \mathcal{B}$, set $C^i_* = \mathsf{msb}_{|M^i_*|}(\overline{C^i_*})$; return $C^i = C^i_1 \| \cdots \| C^i_{m_i} \| C^i_*$ and T^i to $\mathscr{A}$. For $i \in [\widetilde{q}]$, on the i-th decryption query, return $\perp$ to $\mathscr{A}$.

Post-query Phase: Let $L \leftarrow\!\!\$\ \mathcal{L}$ and Π, a permutation over $\mathcal{B}$, be undefined on all blocks, i.e $\mathsf{domain}(\Pi) = \mathsf{range}(\Pi) = \emptyset$.

Step 1: Extending the encryption query-response tuple. Set the following values:

- for $i \in [q]$, $j \in [m^i]$ set $X^i_j = M^i_j \oplus \lambda_1(L, N^i, j)$ and $Y^i_j = C^i_j \oplus \lambda_1(L, N^i, j)$;
- for $i \in \mathcal{I}$, set $X^i_* = \lambda_2(L, N^i, m^i)$ and $Y^i_* = \overline{C^i_*} \oplus \mathsf{pad}(M^i_*)$;
- for $i \in [q] \setminus \mathcal{I}$, set $M^i_{\oplus} = \oplus_{j=1}^{m^i} M^i_j$ and $X^i_{\oplus} = M^i_{\oplus} \oplus \lambda_3(L, N^i, m^i)$;
- for $i \in \mathcal{I}$, set $M^i_{\oplus} = \oplus_{j=1}^{m^i} M^i_j \oplus M^i_*$ and $X^i_{\oplus} = M^i_{\oplus} \oplus \lambda_4(L, N^i, m^i)$;
- for $i \in [q]$, $j \in [a^i]$ set $U^i_j = A^i_j \oplus \lambda_5(L, N^i, j)$;
- for $i \in \mathcal{J}$, set $U^i_* = \mathsf{pad}(A^i_*) \oplus \lambda_6(L, N^i, a^i)$;
- set **bad** $= 1$, if I or O (the partial multiset containing Y^i_j values for $i \in [q]$ and $j \in [m^i] \cup \{*\}$) contains a non-trivial colliding pair (pair of duplicate elements). Note that all colliding pairs are non-trivial at this stage;
- for all $i \in [q]$, $j \in [m^i]$, fix $\Pi(X^i_j) = Y^i_j$, and for all $i \in \mathcal{I}$, fix $\Pi(X^i_*) = Y^i_*$;
- for all $i \in [q]$, $j \in [a^i]$, fix $\Pi(U^i_j) = V^i_j \leftarrow\!\!\$\ \mathcal{B} \setminus \mathsf{range}(\Pi)$, and for all $i \in \mathcal{J}$, fix fix $\Pi(U^i_*) = V^i_* \leftarrow\!\!\$\ \mathcal{B} \setminus \mathsf{range}(\Pi)$;
- for $i \in [q] \setminus \mathcal{J}$, set $S^i = \oplus_{j=1}^{a^i} V^i_j$;
- for $i \in \mathcal{J}$, set $S^i = \oplus_{j=1}^{a^i} V^i_j \oplus V^i_*$;
- for $i \in [q]$ set $Y^i_{\oplus} = S^i \oplus T^i$.
- set **bad** $= 1$, if O (the complete multiset) contains a non-trivial colliding pair (pair of duplicate elements).
- for $i \in [q]$, fix $\Pi(X^i_{\oplus}) = Y^i_{\oplus}$.

Note that the partial sampling of Π remains permutation-compatible in step 1 as long as **bad** $= 0$.

0018-9448 (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/TIT.2019.2925613, IEEE Transactions on Information Theory

Step 2: Extending the decryption query-response tuple. Set the following values:

- at the start all intermediate variables are set as undefined.
- for $i \in [\widetilde{q}]$, $j \in [\widetilde{m}^i]$, set $\widetilde{Y}^i_j = \widetilde{C}^i_j \oplus \lambda_1(L, \widetilde{N}^i, j)$; if $i \in \widetilde{\mathcal{I}}$, set $\widetilde{X}^i_* = \lambda_2(L, \widetilde{N}^i, \widetilde{m}^i)$.
- for $i \in [\widetilde{q}]$, $j \in [\widetilde{m}^i]$, if $\widetilde{Y}^i_j \in \mathsf{range}(\Pi)$ then set $\widetilde{X}^i_j = \Pi^-(\widetilde{Y}^i_j)$ and $\widetilde{M}^i_j = \widetilde{X}^i_j \oplus \lambda_1(L, \widetilde{N}^i, j)$; if $i \in \widetilde{\mathcal{I}}$ and $\widetilde{X}^i_* \in \mathsf{domain}(\Pi)$, then set $\widetilde{Y}^i_* = \Pi(\widetilde{X}^i_*)$ and $\widetilde{M}^i_* = \mathsf{msb}_{|\widetilde{C}^i_*|}(\widetilde{Y}^i_*) \oplus \widetilde{C}^i_*$; if $\widetilde{M}^i$ is defined then set $\widetilde{X}^i_\oplus = \widetilde{M}^i_\oplus \oplus \lambda_3(L, \widetilde{N}^i, \widetilde{m}^i)$.
- for $i \in [\widetilde{q}]$, $j \in [\widetilde{a}^i]$, set $\widetilde{U}^i_j = \widetilde{A}^i_j \oplus \lambda_5(L, \widetilde{N}^i, j)$; if $i \in \widetilde{\mathcal{J}}$, set $\widetilde{U}^i_* = \lambda_6(L, \widetilde{N}^i, \widetilde{a}^i)$.
- for $i \in [\widetilde{q}]$, $j \in [\widetilde{a}^i]$, if $\widetilde{U}^i_j \in \mathsf{domain}(\Pi)$ then set $\widetilde{V}^i_j = \Pi(\widetilde{U}^i_j)$; if $\widetilde{V}^i$ is defined then set $\widetilde{Y}^i_\oplus = \widetilde{S}^i \oplus \widetilde{T}^i$.
- At this stage all those intermediate variables, which can be derived through the MGF key L, and adversary's query, are completely determined. Also, some other variables are trivially derived due to the extended encryption transcript.
- set **bad** $= 1$, if there exists $i \in [\widetilde{q}]$, $j \in [\widetilde{m}^i] \cup \{*, \oplus\}$, and one of the following is true:
  - there exists $i' \in [q]$, $j' \in [m^{i'}] \cup \{*, \oplus\}$ such that $(\widetilde{N}^i, \widetilde{C}^i_j) \neq (N^{i'}, C^{i'}_{j'})$ and $\widetilde{Y}^i_j = Y^{i'}_{j'}$.
  - there exists $i' \in [q]$, $j' \in [a^{i'}] \cup \{*\}$ such that $\widetilde{Y}^i_j = V^{i'}_{j'}$.
- set **bad** $= 1$, if there exists $i \in [\widetilde{q}]$, $j \in [\widetilde{a}^i] \cup \{*\}$, and one of the following is true:
  - there exists $i' \in [q]$, $j' \in [m^{i'}] \cup \{*, \oplus\}$ such that $\widetilde{U}^i_j = X^{i'}_{j'}$.
  - there exists $i' \in [q]$, $j' \in [a^{i'}] \cup \{*\}$ such that $(\widetilde{N}^i, \widetilde{A}^i_j) \neq (N^{i'}, A^{i'}_{j'})$ and $\widetilde{U}^i_j = U^{i'}_{j'}$.

Extend the transcript ω by including all the internal variables computed thus far and return the extended transcript to $\mathscr{A}$.

Identifying the bad transcripts. Let Ω denote the set of all realizable transcripts. We say that $\omega \in \Omega$ is a `Bad` transcript if it causes **bad** $= 1$. In other words, we say that ω is `Bad` if one of the following events occur:

`EEcoll`:

- $\exists\ (i,j) \in [q] \times [m^i] \cup \{*, \oplus\},\ (i',j') \in [q] \times [m^{i'}] \cup \{*, \oplus\}\ :\ (i,j) \neq (i',j') \wedge X^i_j = X^{i'}_{j'}$.
- $\exists\ (i,j) \in [q] \times [a^i] \cup \{*\},\ (i',j') \in [q] \times [a^{i'}] \cup \{*\}\ :\ (i,j) \neq (i',j') \wedge U^i_j = U^{i'}_{j'}$.
- $\exists\ (i,j) \in [q] \times [m^i] \cup \{*, \oplus\},\ (i',j') \in [q] \times [a^{i'}] \cup \{*\}\ :\ X^i_j = U^{i'}_{j'}$.
- $\exists\ (i,j) \in [q] \times [m^i] \cup \{*\},\ (i',j') \in [q] \times [m^{i'}] \cup \{*\}\ :\ (i,j) \neq (i',j') \wedge Y^i_j = Y^{i'}_{j'}$.

`ETcoll`:

- $\exists\ i \in [q],\ (i',j') \in [q] \times [m^{i'}] \cup \{*\}\ :\ Y^i_\oplus = Y^{i'}_{j'}$.
- $\exists\ i \in [q],\ i' \in [q] \setminus \{i\}\ :\ Y^i_\oplus = Y^{i'}_\oplus$.
- $\exists\ i \in [q],\ (i',j') \in [q] \times [a^{i'}] \cup \{*\}\ :\ Y^i_\oplus = V^{i'}_{j'}$.

`DEcoll`:

- $\exists\ (i,j) \in [\widetilde{q}] \times [\widetilde{m}^i] \cup \{*\},\ (i',j') \in [q] \times [m^{i'}] \cup \{*, \oplus\}\ :\ (N^i, C^i_j) \neq (N^{i'}, C^{i'}_{j'}) \wedge \widetilde{Y}^i_j = Y^{i'}_{j'}$.
- $\exists\ (i,j) \in [\widetilde{q}] \times [\widetilde{m}^i] \cup \{*\},\ (i',j')i' \in [q] \times [a^{i'}] \cup \{*\}\ :\ \widetilde{Y}^i_j = V^{i'}_{j'}$.
- $\exists\ (i,j) \in [\widetilde{q}] \times [\widetilde{a}^i] \cup \{*\},\ (i',j') \in [q] \times [a^{i'}] \cup \{*\}\ :\ (\widetilde{N}^i, \widetilde{A}^i_j) \neq (N^{i'}, A^{i'}_{j'}) \wedge \widetilde{U}^i_j = U^{i'}_{j'}$.
- $\exists\ (i,j) \in [\widetilde{q}] \times [\widetilde{a}^i] \cup \{*\},\ (i',j') \in [q] \times [m^{i'}] \cup \{*, \oplus\}\ :\ \widetilde{U}^i_j = X^{i'}_{j'}$.

`DEFcoll`:

- $\exists\ i \in [\widetilde{q}],\ i' \in [q]\ :\ (\widetilde{N}^i, \widetilde{A}^i, \widetilde{T}^i) = (N^{i'}, A^{i'}, T^{i'}) \wedge \widetilde{C}^i = C^{i'}_{1..\widetilde{m}^i} \wedge \widetilde{X}^i_\oplus \in I$.

`DDFcoll`:

- $\exists\ (i,j) \in [\widetilde{q}] \times [\widetilde{m}^i] \cup \{*\},\ i' \in [q]\ :\ (\widetilde{N}^i, \widetilde{A}^i, \widetilde{T}^i) = (N^{i'}, A^{i'}, T^{i'}) \wedge \widetilde{C}^i = C^{i'}_{1..\widetilde{m}^i} \wedge \widetilde{X}^i_\oplus = \widetilde{X}^i_j$.
- $\exists\ (i,j) \in [\widetilde{q}] \times [\widetilde{a}^i] \cup \{*\},\ i' \in [q]\ :\ (\widetilde{N}^i, \widetilde{A}^i, \widetilde{T}^i) = (N^{i'}, A^{i'}, T^{i'}) \wedge \widetilde{C}^i = C^{i'}_{1..\widetilde{m}^i} \wedge \widetilde{X}^i_\oplus = \widetilde{U}^i_j$.
- $\exists\ (i,j) \in [\widetilde{q}] \times [m^i],\ j' \in [\widetilde{m}^i] \setminus \{j\},\ i' \in [q]\ :\ \widetilde{N}^i = N^{i'} \wedge \widetilde{C}^i_j \neq C^{i'}_{j'} \wedge \widetilde{Y}^i_j = \widetilde{Y}^i_{j'}$.
- $\exists\ (i,j) \in [\widetilde{q}] \times [\widetilde{m}^i] \cup \{*, \oplus\}\ :\ \widetilde{Y}^i_1 = \widetilde{Y}^i_j$.

Let $\Omega_{\mathtt{Bad}}$ be the set of all transcripts which are `Bad` and $\Omega_{\mathtt{Good}} = \Omega \setminus \Omega_{\mathtt{Bad}}$. We bound $\mathsf{ip}_{\mathcal{O}_0}[\Omega_{\mathtt{Bad}}]$ to $O(\sigma^2 2^{-n} + r\sigma\epsilon + q\sigma 2^{-n} + q^2 2^{-n} + \widetilde{\sigma}\sigma 2^{-n} + \widetilde{\sigma} r\epsilon + \widetilde{\sigma} 2^{-n} + \widetilde{q}\epsilon)$ in Lemma 6.

Lemma 6.

$$\mathsf{ip}_{\mathcal{O}_0}[\Omega_{\mathtt{Bad}}] \leq \frac{\sigma^2}{2^{n+1}} + \frac{q\sigma + 0.5q^2}{2^n - \sigma} + \frac{\widetilde{\sigma}(\sigma+4)}{2^n} + (\sigma + 3\widetilde{\sigma})r\epsilon + \widetilde{q}\epsilon.$$

Proof. We bound the probability of getting a bad transcript as follows:

$$\begin{aligned}
\mathsf{ip}_{\mathcal{O}_0}[\Omega_{\mathtt{Bad}}] &= \Pr[\mathtt{EEcoll} \vee \mathtt{ETcoll} \vee \mathtt{DEcoll} \vee \mathtt{DEFcoll} \vee \mathtt{DDFcoll}] \\
&\leq \Pr[\mathtt{EEcoll}] + \Pr[\mathtt{ETcoll} \mid \neg\mathtt{EEcoll}] \\
&\quad + \Pr[\mathtt{DEcoll} | \neg(\mathtt{EEcoll} \vee \mathtt{ETcoll})] \\
&\quad + \Pr[\mathtt{DEFcoll} | \neg(\mathtt{EEcoll} \vee \mathtt{ETcoll} \vee \mathtt{DEcoll})] \\
&\quad + \Pr[\mathtt{DDFcoll} | \neg(\mathtt{EEcoll} \vee \mathtt{ETcoll} \vee \mathtt{DEcoll} \vee \mathtt{DEFcoll})]
\end{aligned}$$

BOUNDING $\Pr[\mathtt{EEcoll}]$. Let P and Q be a colliding pair. We bound the probability in two cases:

Case 1: The colliding pair is a non-local pair. As λ is a LIXU hash function, we can bound the probability of this case by 2^{-n}. Further we have at most $\sigma(\sigma-1)/2$ many such pairs. Hence

$$\Pr[\mathtt{EEcoll} \ \wedge\ \text{Case 1}] \leq \frac{\sigma^2}{2^{n+1}}.$$

Case 2: The colliding pair is a local pair. This means that the colliding pair must belong to a single message or associated data. We can have two cases: (2.1) The colliding pair belongs to the i-th message; (2.2) The colliding pair belongs to the i-th associated data. Without loss of generality we assume case

0018-9448 (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/TIT.2019.2925613, IEEE Transactions on Information Theory

(2.1). Let $n^i = \lfloor m^i/r \rfloor$ and $r^i = m^i - n^i r$. As λ is an (ϵ, r)-LIXU hash function, we can bound the probability of a fixed pair by ϵ. Hence

$$\begin{aligned}\Pr\left[\texttt{EEcoll} \ \wedge \ \text{Case 2.1}\right] &\leq \sum_{i=1}^{q}\sum_{j=1}^{n^i}\binom{r}{2}\epsilon + \sum_{i=1}^{q}\binom{r^i}{2}\epsilon \\ &\overset{[a]}{\leq} r'\epsilon\sum_{i=1}^{q} r\cdot n^i + \sum_{i=1}^{q}\binom{r^i}{2}\epsilon \\ &\overset{[b]}{\leq} r'\mu\epsilon - \left(r'\epsilon\sum_{i=1}^{q} r^i - \sum_{i=1}^{q}\binom{r^i}{2}\epsilon\right) \\ &\overset{[c]}{\leq} r'\mu\epsilon.\end{aligned}$$

Here $r' = (r-1)/2$; $[a]$ to $[b]$ follows from $rn^i = m^i - r^i$ and $\sum m^i = \mu$; $[b]$ to $[c]$ follows from the fact that $(r-1)r^i \geq r^i(r^i-1)$ for all i. Similarly the probability in case (2.2) can be bounded to $r'\nu\epsilon$. Hence the total probability in case 2 is bounded by $r'\sigma\epsilon$.

Combining case 1 and 2, we have

$$\Pr\left[\texttt{EEcoll}\right] \leq \frac{\sigma^2}{2^{n+1}} + r\sigma\epsilon.$$

BOUNDING $\Pr\left[\texttt{ETcoll} \mid \neg\texttt{EEcoll}\right]$. This event bounds the probability that the checksum output block collides with some previous output block given that no other input/output collision occurred. At this instant at most $\sigma = \mu + \nu$ many points are defined for Π, whence we can have at most $q\sigma + q^2/2$ many possible colliding pairs. So we have,

$$\Pr\left[\texttt{ETcoll} \mid \neg\texttt{EEcoll}\right] \leq \frac{q\sigma}{2^n - \sigma} + \frac{q^2}{2(2^n-\sigma)}.$$

BOUNDING $\Pr\left[\texttt{DEcoll}|\neg(\texttt{EEcoll} \vee \texttt{ETcoll})\right]$. This event bounds the probability that some intermediate output ($\widetilde{Y}^i_j$) or input ($\widetilde{U}^i_j$) input block non-trivially belongs in O or I, respectively. First, suppose $\widetilde{Y}^i_j$ belongs to O for some $(i,j) \in [\widetilde{q}] \times [\widetilde{m}^i]$. Now as in case of EEcoll above: i) the colliding encryption block can either be a non-local pair of (i,j), in which case we bound the probability to at most $\sigma/2^n$; or ii) the colliding encryption block can be a local pair of (i,j), in which case we bound the probability to at most $r\epsilon$. Since there are at most $\widetilde{\mu}$ many decryption ciphertext blocks, the probability that there exists $(i,j) \in [\widetilde{q}] \times [\widetilde{m}^i]$, such that $\widetilde{Y}^i_j$ belongs in O non-trivially is bounded by $\widetilde{\mu}\sigma/2^n + \widetilde{\mu}r\epsilon$. Similarly, one can bound the probability that there exists $(i,j) \in [\widetilde{q}] \times [\widetilde{a}^i]$, such that $\widetilde{U}^i_j$ belongs in I non-trivially to at most $\widetilde{\nu}\sigma/2^n + \widetilde{\nu}r\epsilon$. Combining the two cases, we have

$$\Pr\left[\texttt{DEcoll}|\neg(\texttt{EEcoll} \vee \texttt{ETcoll})\right] \leq \frac{\widetilde{\sigma}\sigma}{2^n} + \widetilde{\sigma}r\epsilon.$$

BOUNDING $\Pr\left[\texttt{DEFcoll}|\neg(\texttt{EEcoll} \vee \texttt{ETcoll} \vee \texttt{DEcoll})\right]$. This case bounds the probability that some decryption checksum input block collides with some encryption input block with a restriction that the decryption input blocks must all be defined. In this case we have $\widetilde{X}^i_{\oplus} = \sum_{j=1}^{\widetilde{m}^i} \widetilde{M}^i_j \oplus \lambda_3(L, \widetilde{N}, \widetilde{m}^i)$. Now we can have two cases: i) the checksum block collides with $X^{i'}_{\oplus}$, which can be generously bounded by ϵ (assuming $(\widetilde{N}^i, \widetilde{m}^i)$ and $(N^{i'}, m^{i'})$ are local pair); and ii) the checksum block collides with any other $X^{i'}_{j'}$, in which case we bound the probability by 2^{-n} using the 1-universal property of λ. In summary, we have

$$\Pr\left[\texttt{DEFcoll}|\neg(\texttt{EEcoll} \vee \texttt{ETcoll} \vee \texttt{DEcoll})\right] \leq \frac{\widetilde{q}}{2^n} + \widetilde{q}\epsilon.$$

BOUNDING $\Pr\left[\texttt{DDFcoll}|\neg(\texttt{EEcoll} \vee \texttt{ETcoll} \vee \texttt{DEcoll} \vee \texttt{DEFcoll})\right]$. This case simply bound the probability that there is no collisions within a decryption query. Let us fix a decryption query index $i \in \widetilde{q}$. Then the first case is bounded by at most $(\widetilde{m}^i 2^{-n}$; the second case is bounded by at most $\widetilde{a}^i 2^{-n}$; the third case is bounded by at most $(\widetilde{m}^i - r)2^{-n} + r\epsilon$; and the fourth case is bounded by at most $(\widetilde{m}^i - r)2^{-n} + r\epsilon$. On combining the four cases, we have

$$\Pr\left[\texttt{DDFcoll}|\neg(\texttt{EEcoll} \vee \texttt{ETcoll} \vee \texttt{DEcoll} \vee \texttt{DEFcoll})\right] \leq \frac{3\widetilde{\sigma}}{2^n} + 2\widetilde{q}r\epsilon.$$

The result follows by combining all the individual bounds above, and simplifying using $\widetilde{q} \leq \widetilde{\sigma}$. □

Ratio of interpolation probabilities. Fix an arbitrary transcript $\omega \in \Omega_{\texttt{Good}}$. In Lemma 7, we show that the ratio $\mathsf{ip}_{\mathcal{O}_1}[\omega]/\mathsf{ip}_{\mathcal{O}_0}[\omega]$ is at least $(1 - O(\widetilde{q}2^{-n}))$.

Lemma 7. *For any $\omega \in \Omega_{\texttt{Good}}$, we have*

$$\frac{\mathsf{ip}_{\mathcal{O}_1}[\omega]}{\mathsf{ip}_{\mathcal{O}_0}[\omega]} \geq \left(1 - \frac{\widetilde{q}}{2^n - \sigma}\right).$$

Proof. In $\mathcal{O}_0$, for encryption phase: first $\mu + q$ ciphertext and tag outputs are sampled in with replacement fashion, followed by the sampling of ν output blocks in without replacement manner from a subset of $\mathcal{B}$ of size $(2^n - \mu)$ (Π is already defined on μ many input-output blocks); for decryption phase the oracle always returns $\perp$. Hence

$$\mathsf{ip}_{\mathcal{O}_0}[\omega] = \frac{1}{(2^n)^{\mu+q}\cdot(2^n-\mu)^{\underline{\nu}}}. \tag{10}$$

In $\mathcal{O}_1$, for any ω we denote the encryption tuples by ω_e and the decryption tuples by ω_d. Then we have

$$\begin{aligned}\mathsf{ip}_{\mathcal{O}_1}[\omega] &= \Pr_{\mathcal{O}_1}[\omega_e, \omega_d] \\ &= \Pr_{\mathcal{O}_1}[\omega_e] \cdot \Pr_{\mathcal{O}_1}[\omega_d \mid \omega_e] \\ &= \Pr_{\mathcal{O}_1}[\omega_e] \cdot \left(1 - \Pr_{\mathcal{O}_1}[\neg\omega_d \mid \omega_e]\right) \end{aligned} \tag{11}$$

where $\omega_d = (\widetilde{N}^i, \widetilde{A}^i, \widetilde{C}^i, \widetilde{T}^i, \perp)_{i\in[\widetilde{q}]}$, as the ideal oracle always returns $\perp$ on decryption queries. Note that we have slightly abused the notation to use $\neg\omega_d$ as the event that: for some $i \in [\widetilde{q}]$ the i-th decryption query successfully decrypts.

For encryption tuples, exactly $\mu+\nu+q$ many calls are made to Π: one for each of the μ message blocks, one for each of the ν associated data blocks; and one for each of the q tags. As $\sigma = \mu + \nu$ we have

$$\Pr_{\mathcal{O}_1}[\omega_e] = \frac{1}{(2^n)^{\underline{\sigma+q}}}. \tag{12}$$

Now we upper bound the probability of $\neg\omega_d$. It is enough to bound the probability for a fixed index $i \in [\widetilde{q}]$ corresponding to

0018-9448 (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/TIT.2019.2925613, IEEE Transactions on Information Theory

a successful decryption query. By union bound, the probability of $\neg\omega_d$ is at most $\widetilde{q}$ times more than the single forgery probability. Without loss of generality assume that $i \in [\widetilde{q}] \setminus \widetilde{\mathcal{I}}$ and $i \in [\widetilde{q}] \setminus \widetilde{\mathcal{J}}$. Note that the adversary succeeds in forgery if the adversary somehow makes the i-th query in such a way that the following equation holds.

$$\Pi(\widetilde{X}^i_{\oplus}) \oplus \widetilde{Y}^i_{\oplus} = 0 \tag{13}$$

Now based on the i-th decryption query we can have different cases:

Case 1: $\exists \boldsymbol{j} \in [\boldsymbol{q}],\ \widetilde{\boldsymbol{N}}^{\boldsymbol{i}} = \boldsymbol{N}^{\boldsymbol{j}}$. Based on $\widetilde{C}^i$ we can have two subcases:

Subcase 1.1: $\widetilde{\boldsymbol{C}}^{\boldsymbol{i}} = \boldsymbol{C}^{\boldsymbol{i}'}_{\boldsymbol{1..\widetilde{m}^i}}$. In this case $\widetilde{M}^i_{\oplus}$ is pre-determined as $\oplus_{k=1}^{\widetilde{c}^i} M^{i'}_k$. Suppose $\widetilde{m}^i = m^{i'}$, i.e. the two ciphertexts are exactly the same. Then we must have some j such that $\widetilde{A}^i_j \neq A^{i'}_j$, otherwise i-th decryption query is a duplicate of i'-th encryption query. Since the transcript is good, $\widetilde{U}^i_j$ is fresh. So, by exploiting the conditional randomness of Π we can bound the probability to at most $1/(2^n - \sigma)$. Now suppose $\widetilde{C}^i$ is a proper prefix of $C^{i'}$, then we must have a fresh $\widetilde{X}^i_{\oplus}$ (as the transcript is good), whence we can again bound the probability to at most $1/(2^n - \sigma)$.

The two cases discussed above are mutually exclusive, hence the probability that the adversary forges in subcase 2.1 is at most $1/(2^n - \sigma)$.

Subcase 1.2: $\widetilde{\boldsymbol{C}}^{\boldsymbol{i}} \neq \boldsymbol{C}^{\boldsymbol{i}'}_{\boldsymbol{1..\widetilde{c}^i}}$. In this case there exists at least one $k \in [\widetilde{c}^i]$ such that $\widetilde{C}^i_k \neq C^{i'}_k$. We consider first such block $\widetilde{C}^i_k$. Since the transcript is good, we must have a fresh $\widetilde{Y}^i_k$. Observe that we can rewrite Eq. (13) as

$$\Pi^-(\widetilde{Y}^i_k) = \Pi^-(\widetilde{Y}^i_{\oplus}) \oplus \delta \oplus \lambda_3(L, \widetilde{N}^i, \widetilde{m}^i),$$

where δ denotes the checksum of all the blocks of i-th query, except the k-th block. Again, by conditioning on all points except $\widetilde{Y}^i_k$, we bound the probability of this case by $1/(2^n - \sigma)$.

Case 2: $\forall \boldsymbol{j} \in [\boldsymbol{q}],\ \widetilde{\boldsymbol{N}}^{\boldsymbol{i}} \neq \boldsymbol{N}^{\boldsymbol{j}}$. In this case $\widetilde{N}^i$ has not been used for mask generation till now. Without loss of generality, we assume that $\widetilde{m}^i \geq 1$. Since the transcript is good, we must have a fresh $\widetilde{Y}^i_1$. Now using a similar line of argument as in sub case 1.2 we bound the probability of this event by $1/(2^n - \sigma)$.

Note that case 1.1, case 1.2 and case 2 are all mutually exclusive so we can take the maximum over all three. As $\sigma \geq a^i + a^j$ we have

$$\Pr_{\mathcal{O}_1}[\neg\omega_d \mid \omega_e] \leq \sum_{i=1}^{\widetilde{q}} \frac{1}{2^n - \sigma} \leq \frac{\widetilde{q}}{2^n - \sigma} \tag{14}$$

On substituting (12) and (14) in (11) and dividing the result by (10) we have

$$\begin{aligned}
\frac{\mathsf{ip}_{\mathcal{O}_1}[\omega]}{\mathsf{ip}_{\mathcal{O}_0}[\omega]} &\geq \frac{(2^n)^{\underline{\mu+q}} \cdot (2^n - \mu)^{\underline{\nu}}}{(2^n)^{\underline{\sigma+q}}} \left(1 - \frac{\widetilde{q}}{2^n - \sigma}\right) \\
&\geq \frac{(2^n)^{\underline{q}}}{(2^n - \sigma)^{\underline{q}}} \left(1 - \frac{\widetilde{q}}{2^n - \sigma}\right) \\
&\geq \left(1 - \frac{\widetilde{q}}{2^n - \sigma}\right).
\end{aligned}$$

□

Using Lemma 6, Lemma 7 and Theorem 2 we get the desired result.

D. Proof of Lemma 4 and Lemma 5

The proofs of these lemmata are quite similar. We give the proof for Lemma 4, while the proof for Lemma 5 can be similarly obtained.

Fix distinct $(N, i), (N', i') \in \mathcal{N} \times \mathbb{F}_{2^8}$ and $\delta \in \mathcal{B}$. Now we have two cases:

Case 1: (N, i) and (N', i') are local pairs. We know that $N = N'$ and $i \neq i'$. Hence,

$$\begin{aligned}
&\Pr\left[\mathsf{A1}_\Gamma(N, i) \oplus \mathsf{A1}_\Gamma(N', i') = \delta\right] \\
&\quad = \Pr\left[\mathsf{aes1}_{\Gamma(N)}(i) \oplus \mathsf{aes1}_{\Gamma(N)}(i') = \delta\right] \\
&\quad \leq \frac{1}{2^{96}},
\end{aligned}$$

where the last inequality follows from Lemma 2.

Case 2: (N, i) and (N', i') are not local pairs. We know that $N \neq N'$. Hence,

$$\begin{aligned}
&\Pr\left[\mathsf{A1}_\Gamma(N, i) \oplus \mathsf{A1}_\Gamma(N', i') = \delta\right] \\
&\quad = \Pr\left[\mathsf{aes1}_{\Gamma(N)}(i) \oplus \mathsf{aes1}_{\Gamma(N')}(i') = \delta\right] \\
&\quad \leq \Pr\left[\Gamma(N) = \text{1-}\mathtt{AESRD}^{-1}(\mathsf{aes1}_{\Gamma(N')}(i') \oplus \delta) \oplus i\right] \\
&\quad \leq \tfrac{1}{2^{128}}.
\end{aligned}$$

The result follows from the bounds in case 1 and case 2.

VII. Software Performance

In this section, we present software implementation of GOCB instantiated with AES-128 as the underlying block cipher, and A1 and A2 as the underlying MGFs, and compare the performance of the proposed designs against the standard OCB3-AES-128. We do the benchmarking in two cases: i) sequential processing (the conventional implementation), and ii) random read access (introduced in this paper).

Implementation — We reuse the optimized C code of OCB3 [9] by Krovetz. Further, we introduce minimal changes, as required, over the OCB3 code to generate the C code for GOCB[AES-128,A1], and GOCB[AES-128,A2]. We also reuse the time measurement mechanism as used in [9]. In summary, we utilized the hardware support on our benchmarking platform for Intel's SSE4 vector instructions [59], Intel's `pclmulqdq` instruction for multiplication over $\mathbb{F}_{2^{128}}$, and Intel's AES-NI library [33] for operations involving AES-128 round functions. We made some simplifying assumptions for the implementations:

- message length is treated to be equal to the length of the associated data, and
- all messages and associated data are assumed to have complete blocks.

Platform — We performed the benchmarking on:

- Intel Core i7-6500U "Skylake" (2.5 GHz with 64 KB L1 cache, 256 KB L2 cache, 4 MB L3 cache) with Ubuntu 16.04 LTS (kernel version 4.4),

0018-9448 (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/TIT.2019.2925613, IEEE Transactions on Information Theory

which supports Intel's SSE4 vector instructions [59], standard AES-NI instructions [33] and `pclmuldq` [60], the carry-less multiplication instruction.

SETUP — The tests for benchmark were compiled using `gcc` version 5.4, with optimization flag `-O3`, and instruction set architecture flag `-march=native` (as instructed in [9]). All the tests were run on isolated core, after turning off processor frequency scaling and power management options such as Intel's Turbo Boost and Hyper-Threading technologies. The reference baseline performance for AES-128, using AES-NI instruction set, is presented in Table II. All performance figures presented are throughputs, in units of cycles-per-byte (cpb).

TABLE II: Baseline throughput (cpb) of AES using AES-NI.

Microarchitecture	Encryption (cpb)
Intel Core i7-6500U "Skylake"	0.66

TESTING METHODOLOGY — Following [9], we measure the cpb value for every message length from 1 to 1024 bytes, as well as 1500 and 4096 bytes. To avoid any discrepancies arising due to memory subsystem, we execute warm up runs so that the all code and data is in the cache before benchmarking begins. To capture the average cost of encryption, we use the `rdtsc` instruction — a time-stamp counter available on Skylake processors — to time encryption of the same message: i) 64 times for sequential processing; ii) 1536 times for random access. Note that we average over a larger number of repetitions in random access case to get an appreciable measurement for a single block encryption. The median of 15 such averaged values is reported as the number of cycles. The cycles per byte value is obtained by dividing the median value by the length of the corresponding message.

TABLE III: Performance comparison between sequential implementations of OCB3-AES-128, GOCB[AES-128,A1], and GOCB[AES-128,A2]. The performance figures presented are throughputs, in units of cycles-per-byte (cpb).

Length	OCB3-AES-128	GOCB[AES-128,A1]	GOCB[AES-128,A2]
128 bytes	1.48	2.11	2.15
256 bytes	1.15	1.41	1.45
512 bytes	0.85	1.07	1.10
1024 bytes	0.75	0.90	0.93
4096 bytes	0.68	0.77	0.81

TABLE IV: Summary of IPI values for OCB3-AES-128, GOCB[AES-128,A1], and GOCB[AES-128,A2].

	OCB3-AES-128	GOCB[AES-128,A1]	GOCB[AES-128,A2]
IPI:	1.05	1.24	1.28

SEQUENTIAL PROCESSING: RESULTS AND DISCUSSION — Table III summarizes the cpb values for OCB3-AES-128, GOCB[AES-128,A1], and GOCB[AES-128,A2], for moderate to large message lengths, in sequential processing. Following [9], we also present IPI (Internet Performance Index) values — a weighted average cpb vlaue for usual message lengths observed over internet — in Table IV, for OCB3-AES-128, GOCB[AES-128,A1], and GOCB[AES-128,A2]. As per our experiments, OCB3 with 10 rounds of AES-128 is approximately, 9% and 10% faster than OCB3 with 11 and 12 rounds, respectively, of AES-128. One may think that GOCB[AES-128,A1] and GOCB[AES-128,A2] should have similar degradation in performance. But the IPI clearly indicates that the A1 and A2 variants are much more slower, approximately 18% and 22%, respectively, than OCB3 over usual internet data. This is mainly due to the lower masking (just after encryption), which is the contentious instruction that breaks the pipelining benefits.

However, the important point to note is that for large messages (i.e. ≥ 4096 bytes), which is our main area of focus, the cpb values for GOCB[AES-128,A1] and GOCB[AES-128,A2] tend towards the cpb values for OCB3-AES-128 with 11 and 12 rounds, respectively. Specifically, we have observed that OCB3-AES-128 with 11 and 12 rounds have 0.74 and 0.81 cpbs, respectively, for 4096 bytes message. This is quite close to the cpb values achieved by GOCB[AES-128,A1] and GOCB[AES-128,A2], i.e. 0.77 and 0.81, respectively. Beyond this point the cpb values for GOCB[AES-128,A1] and GOCB[AES-128,A2] are expected to saturate and follow the cpb values of OCB3-AES-128 with 11 and 12 rounds, respectively. In other words, for messages with length ≥ 4096 bytes, GOCB[AES-128,A1] and GOCB[AES-128,A2] are only $\approx 10\%$ and $\approx 20\%$ slower than OCB3.

TABLE V: Summary of cpb values for OCB3-AES-128, GOCB[AES-128,A1], and GOCB[AES-128,A2] to process a single block of data in random access.

OCB3-AES-128	GOCB[AES-128,A1]	GOCB[AES-128,A2]
3.93	3.12	3.29

RANDOM ACCESS: RESULTS AND DISCUSSION— To the best of our knowledge, optimized 128-bit field multiplication algorithms (using `pclmulqdq`) are very close to 5 rounds of AES (using AES-NI) in terms of performance. Thus, theoretically one would expect that GOCB[AES-128,A1], and GOCB[AES-128,A2] will have significantly better random access performance as compared to OCB3, which requires a 128-bit field multiplication.

In Skylake, 1-round AES call requires 4 cycles (latency), so GOCB[AES-128,A1] and GOCB[AES-128,A2] would require close to 44 and 48 cycles, respectively, for AES round calls. Further, roughly 4 cycles are required to create the initial masking state. So, theoretically GOCB[AES-128,A1] and GOCB[AES-128,A2] are expected to have approximately 3-3.1 and 3.2-3.3 cpb. On the other hand, OCB3 requires close to 60 cycles for field multiplication and the AES call. Apart from this approx. 3 cycles are required for gray code generation. So, OCB3 is expected to have around 3.9-4 cpb. In other words, GOCB[AES-128,A1] and GOCB[AES-128,A2] are expected to be close to 25% and 20%, respectively, better than OCB3 in terms of random access. Note that we have ignored the overhead due to nonce processing. This is reasonable given

0018-9448 (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/TIT.2019.2925613, IEEE Transactions on Information Theory

the assumption that random read/write does not span across different messages.

Table V summarizes the cpb values for OCB3-AES-128, GOCB[AES-128,A1], and GOCB[AES-128,A2], to process an arbitrary single block. Clearly, the experimental data justifies the theoretical speedup described above.

PERFORMANCE IN NON AES-NI ENVIRONMENTS — In environments which lack AES-NI instructions, field multiplication instruction (required for random read in existing OCB designs) is also not available. To the best of our knowledge, optimized implementations of two rounds of AES is expected to be significantly faster than 128-bit field multiplications. So we expect that A1 and A2 based GOCB will be faster than OCB3, in random access, even on non AES-NI environments, such as GPU based applications.

VIII. CONCLUSION

Most parallel message authentication and authenticated encryption schemes use certain masking routines based on the position of a message block and/or a nonce. We explicitly define this idea as the *Mask Generating Function (MGF)*. We used this abstraction of the masking scheme to explore efficient solutions for random read access in OCB variants. Specifically, the MGF based generalization of OCB design, called GOCB, allows us to prove that a relaxed notion of universal hash — Locally-Imperfect XOR Universal Hash (LIXU) — is sufficient to provide adequate security guarantees for masking. We demonstrate secure instantiations of LIXU hash functions in GOCB design using 1-round and 2-round AES, which are expected to make direct computation extremely efficient. This in turn allows for efficient random read access in GOCB. We demonstrate that even though 1-round and 2-round AES do not have enough differential probability (MDP) to qualify as optimal AXU hash functions, they do just fine as MGFs in the GOCB design if we consider the LIXU security notion for masking. The abstraction of LIXU hash functions for mask generation, and the proposed instantiation of LIXU hash using 1-round and 2-round AES in GOCB like construction are expected to have a number of benefits in practice, including — random read access, out-of-sequence encryption/decryption, high parallelism, and support for constant time implementations — to name a few. It will be interesting to see tailor-made applications and instantiations of LIXU hash functions in other practical cryptographic constructions in the future. For instance, it would be interesting to see whether the XTS mode by Rogaway [8] can be modified using LIXU hash function.

REFERENCES

[1] M. Bellare and C. Namprempre, "Authenticated encryption: Relations among notions and analysis of the generic composition paradigm," in *Advances in Cryptology - ASIACRYPT 2000, 6th International Conference on the Theory and Application of Cryptology and Information Security, Kyoto, Japan, December 3-7, 2000, Proceedings*, 2000, pp. 531–545.

[2] C. S. Jutla, "Encryption modes with almost free message integrity," in *Advances in Cryptology - EUROCRYPT 2001, International Conference on the Theory and Application of Cryptographic Techniques, Innsbruck, Austria, May 6-10, 2001, Proceeding*, 2001, pp. 529–544.

[3] P. Rogaway, M. Bellare, and J. Black, "OCB: A block-cipher mode of operation for efficient authenticated encryption," *ACM Trans. Inf. Syst. Secur.*, vol. 6, no. 3, pp. 365–403, 2003.

[4] CAESAR, "Competition for Authenticated Encryption: Security, Applicability, and Robustness," 2014.

[5] A. Chakraborti, A. Chattopadhyay, M. Hassan, and M. Nandi, "Trivia: A fast and secure authenticated encryption scheme," in *Cryptographic Hardware and Embedded Systems - CHES 2015 - 17th International Workshop, Saint-Malo, France, September 13-16, 2015, Proceedings*, 2015, pp. 330–353.

[6] P. Sarkar, "Modes of operations for encryption and authentication using stream ciphers supporting an initialisation vector," *Cryptography and Communications*, vol. 6, no. 3, pp. 189–231, 2014.

[7] G. Bertoni, J. Daemen, M. Peeters, and G. V. Assche, "Duplexing the sponge: Single-pass authenticated encryption and other applications," in *Selected Areas in Cryptography - 18th International Workshop, SAC 2011, Toronto, ON, Canada, August 11-12, 2011, Revised Selected Papers*, 2011, pp. 320–337.

[8] P. Rogaway, "Efficient instantiations of tweakable blockciphers and refinements to modes OCB and PMAC," in *Advances in Cryptology - ASIACRYPT 2004, 10th International Conference on the Theory and Application of Cryptology and Information Security, Jeju Island, Korea, December 5-9, 2004, Proceedings*, 2004, pp. 16–31.

[9] T. Krovetz and P. Rogaway, "The software performance of authenticated-encryption modes," in *Fast Software Encryption - 18th International Workshop, FSE 2011, Lyngby, Denmark, February 13-16, 2011, Revised Selected Papers*, 2011, pp. 306–327.

[10] D. A. McGrew and J. Viega, "The security and performance of the galois/counter mode (GCM) of operation," in *Progress in Cryptology - INDOCRYPT 2004, 5th International Conference on Cryptology in India, Chennai, India, December 20-22, 2004, Proceedings*, 2004, pp. 343–355.

[11] E. Andreeva, A. Bogdanov, A. Luykx, B. Mennink, E. Tischhauser, and K. Yasuda, "Parallelizable and authenticated online ciphers," in *Advances in Cryptology - ASIACRYPT 2013 - 19th International Conference on the Theory and Application of Cryptology and Information Security, Bengaluru, India, December 1-5, 2013, Proceedings, Part I*, 2013, pp. 424–443.

[12] L. Bossuet, N. Datta, C. Mancillas-López, and M. Nandi, "Elmd: A pipelineable authenticated encryption and its hardware implementation," *IEEE Trans. Computers*, vol. 65, no. 11, pp. 3318–3331, 2016.

[13] N. Datta, A. Luykx, B. Mennink, and M. Nandi, "Understanding RUP integrity of COLM," *IACR Trans. Symmetric Cryptol.*, vol. 2017, no. 2, pp. 143–161, 2017.

[14] J. Jean, I. Nikolic, and T. Peyrin, "Tweaks and keys for block ciphers: The TWEAKEY framework," in *Advances in Cryptology - ASIACRYPT 2014 - 20th International Conference on the Theory and Application of Cryptology and Information Security, Kaoshiung, Taiwan, R.O.C., December 7-11, 2014, Proceedings, Part II*, 2014, pp. 274–288.

[15] T. Peyrin and Y. Seurin, "Counter-in-tweak: Authenticated encryption modes for tweakable block ciphers," in *Advances in Cryptology - CRYPTO 2016 - 36th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 14-18, 2016, Proceedings, Part I*, 2016, pp. 33–63.

[16] M. Dworkin, "Recommendation for block cipher modes of operation – methods and techniques," National Institute of Standards and Technology, U. S. Department of Commerce, NIST Special Publication 800-38A, 2001.

[17] D. Whiting, R. Housley, and N. Ferguson, "Counter with CBC-MAC (CCM)," Internet Engineering Task Force, RFC 3610, 2003.

[18] P. Rogaway, "Authenticated-encryption software performance: Comparison of CCM, GCM, and OCB," http://web.cs.ucdavis.edu/~rogaway/ocb/performance/, (Accessed: 2018-06-19).

[19] ——, "How much faster is OCB than CCM or GCM?" http://web.cs.ucdavis.edu/~rogaway/ocb/ocb-faq.htm#performance, (Accessed: 2018-06-19).

[20] J. Postel, "User datagram protocol," Internet Engineering Task Force, RFC 768, 1980.

[21] M. Baugher, D. McGrew, M. Naslund, E. Carrara, and K. Norrman, "The secure real-time transport protocol (SRTP)," *RFC*, vol. 3711, pp. 1–56, 2004.

[22] I. Johansson and M. Westerlund, "Support for reduced-size real-time transport control protocol (RTCP): Opportunities and consequences," *RFC*, vol. 5506, pp. 1–17, 2009.

[23] J. Lennox, "Encryption of header extensions in the secure real-time transport protocol (SRTP)," *RFC*, vol. 6904, pp. 1–15, 2013.

0018-9448 (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/TIT.2019.2925613, IEEE Transactions on Information Theory

[24] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "RTP: A transport protocol for real-time applications," *RFC*, vol. 3550, pp. 1–89, 2003.

[25] J. Black and P. Rogaway, "A block-cipher mode of operation for parallelizable message authentication," *IACR Cryptology ePrint Archive*, pp. 1–27, 2001.

[26] R. Granger, P. Jovanovic, B. Mennink, and S. Neves, "Improved masking for tweakable blockciphers with applications to authenticated encryption," in *Advances in Cryptology - EUROCRYPT 2016 - 35th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Vienna, Austria, May 8-12, 2016, Proceedings, Part I*, 2016, pp. 263–293.

[27] D. Chakraborty and P. Sarkar, "A general construction of tweakable block ciphers and different modes of operations," *IEEE Trans. Information Theory*, vol. 54, no. 5, pp. 1991–2006, 2008.

[28] S. Contini and Y. Yin, "On differential properties of data-dependent rotations and their use in MARS and RC6 (Extended Abstract)," in *Second AES Candidate Conference, March 22-23, 1999, Proceedings*, 2000, pp. 230–239.

[29] K. Minematsu, "A short universal hash function from bit rotation, and applications to blockcipher modes," in *Provable Security - 7th International Conference, ProvSec 2013, Melaka, Malaysia, October 23-25, 2013. Proceedings*, 2013, pp. 221–238.

[30] S. Park, S. H. Sung, S. Lee, and J. Lim, "Improving the upper bound on the maximum differential and the maximum linear hull probability for SPN structures and AES," in *Fast Software Encryption, 10th International Workshop, FSE 2003, Lund, Sweden, February 24-26, 2003, Revised Papers*, 2003, pp. 247–260.

[31] L. Keliher, "Refined analysis of bounds related to linear and differential cryptanalysis for the AES," in *Advanced Encryption Standard - AES, 4th International Conference, AES 2004, Bonn, Germany, May 10-12, 2004, Revised Selected and Invited Papers*, 2004, pp. 42–57.

[32] L. Keliher and J. Sui, "Exact maximum expected differential and linear probability for two-round Advanced Encryption Standard," *IET Information Security*, vol. 1, no. 2, pp. 53–57, 2007.

[33] S. Gueron, "Intel® Advanced Encryption Standard (AES) new instructions set," Intel Corporation, Tech. Rep., 2010.

[34] K. Minematsu, "Improved security analysis of XEX and LRW modes," in *Selected Areas in Cryptography, 13th International Workshop, SAC 2006, Montreal, Canada, August 17-18, 2006 Revised Selected Papers*, 2006, pp. 96–113.

[35] K. Minematsu and Y. Tsunoo, "Provably secure MACs from differentially-uniform permutations and AES-based implementations," in *Fast Software Encryption, 13th International Workshop, FSE 2006, Graz, Austria, March 15-17, 2006, Revised Selected Papers*, 2006, pp. 226–241.

[36] M. Bellare, V. T. Hoang, and S. Keelveedhi, "Cryptography from compression functions: The UCE bridge to the ROM," in *Advances in Cryptology - CRYPTO 2014 - 34th Annual Cryptology Conference, Santa Barbara, CA, USA, August 17-21, 2014, Proceedings, Part I*, 2014, pp. 169–187.

[37] R. Bhaumik and M. Nandi, "Improved security for OCB3," in *Advances in Cryptology - ASIACRYPT 2017 - 23rd International Conference on the Theory and Applications of Cryptology and Information Security, Hong Kong, China, December 3-7, 2017, Proceedings, Part II*, 2017, pp. 638–666.

[38] H. Krawczyk, "Lfsr-based hashing and authentication," in *Advances in Cryptology - CRYPTO '94, 14th Annual International Cryptology Conference, Santa Barbara, California, USA, August 21-25, 1994, Proceedings*, 1994, pp. 129–139.

[39] P. Rogaway, "Bucket hashing and its application to fast message authentication," *J. Cryptology*, vol. 12, no. 2, pp. 91–115, 1999.

[40] K. Minematsu, "Parallelizable rate-1 authenticated encryption from pseudorandom functions," in *Advances in Cryptology - EUROCRYPT 2014 - 33rd Annual International Conference on the Theory and Applications of Cryptographic Techniques, Copenhagen, Denmark, May 11-15, 2014. Proceedings*, 2014, pp. 275–292.

[41] K. Yasuda, "On the full MAC security of a double-piped mode of operation," *IEICE Transactions*, vol. 94-A, no. 1, pp. 84–91, 2011.

[42] S. Halevi and P. Rogaway, "A tweakable enciphering mode," in *Advances in Cryptology - CRYPTO 2003, 23rd Annual International Cryptology Conference, Santa Barbara, California, USA, August 17-21, 2003, Proceedings*, 2003, pp. 482–499.

[43] T. Krovetz and P. Rogaway, "The OCB authenticated-encryption algorithm," *RFC*, vol. 7253, pp. 1–19, 2014.

[44] M. Ågren, M. Hell, T. Johansson, and W. Meier, "Grain-128a: a new version of Grain-128 with optional authentication," *IJWMC*, vol. 5, no. 1, pp. 48–59, 2011.

[45] J. Daemen and V. Rijmen, *The Design of Rijndael: AES - The Advanced Encryption Standard*, ser. Information Security and Cryptography. Springer, 2002.

[46] ——, "Understanding two-round differentials in AES," in *Security and Cryptography for Networks, 5th International Conference, SCN 2006, Maiori, Italy, September 6-8, 2006, Proceedings*, 2006, pp. 78–94.

[47] NIST, "Announcing the ADVANCED ENCRYPTION STANDARD (AES)," National Institute of Standards and Technology, U. S. Department of Commerce, Fedral Information Processing Standards Publication FIPS 197, 2001.

[48] M. Matsui and A. Yamagishi, "A new method for known plaintext attack of FEAL cipher," in *Advances in Cryptology - EUROCRYPT '92, Workshop on the Theory and Application of of Cryptographic Techniques, Balatonfüred, Hungary, May 24-28, 1992, Proceedings*, 1992, pp. 81–91.

[49] E. Biham and A. Shamir, "Differential cryptanalysis of des-like cryptosystems," in *Advances in Cryptology - CRYPTO '90, 10th Annual International Cryptology Conference, Santa Barbara, California, USA, August 11-15, 1990, Proceedings*, 1990, pp. 2–21.

[50] G. Jakimoski and K. P. Subbalakshmi, "On efficient message authentication via block cipher design techniques," in *Advances in Cryptology - ASIACRYPT 2007, 13th International Conference on the Theory and Application of Cryptology and Information Security, Kuching, Malaysia, December 2-6, 2007, Proceedings*, 2007, pp. 232–248.

[51] D. J. Bernstein, "Cache-timing attacks on AES," 2005. [Online]. Available: https://cr.yp.to/antiforgery/cachetiming-20050414.pdf

[52] NIST LwC Standardization Team, "Submission requirements and evaluation criteria for the lightweight cryptography standardization process," http://web.cs.ucdavis.edu/~rogaway/ocb/ocb-faq.htm#performance, 2018, (Accessed: 2019-04-09).

[53] J. Patarin, "Etude des générateurs de permutations pseudo-aléatoires basés sur le schéma du DES," Ph.D. dissertation, Université de Paris, 1991.

[54] ——, "The "Coefficients H" technique," in *Selected Areas in Cryptography, 15th International Workshop, SAC 2008, Sackville, New Brunswick, Canada, August 14-15, Revised Selected Papers*, 2008, pp. 328–345.

[55] B. Mennink and S. Neves, "Encrypted Davies-Meyer and its dual: Towards optimal security using mirror theory," in *Advances in Cryptology - CRYPTO 2017 - 37th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 20-24, 2017, Proceedings, Part III*, 2017, pp. 556–583.

[56] S. Chen and J. P. Steinberger, "Tight security bounds for key-alternating ciphers," in *Advances in Cryptology - EUROCRYPT 2014 - 33rd Annual International Conference on the Theory and Applications of Cryptographic Techniques, Copenhagen, Denmark, May 11-15, 2014. Proceedings*, 2014, pp. 327–350.

[57] D. J. Bernstein, "How to stretch random functions: The security of protected counter sums," *J. Cryptology*, vol. 12, no. 3, pp. 185–192, 1999.

[58] M. Nandi, "A simple and unified method of proving indistinguishability," in *Progress in Cryptology - INDOCRYPT 2006, 7th International Conference on Cryptology in India, Kolkata, India, December 11-13, 2006, Proceedings*, 2006, pp. 317–334.

[59] *Intel® SSE4 (AES) Programming Reference*, Intel Corporation, 2007.

[60] S. Gueron, "Intel® carry-less multiplication instruction and its usage for computing the GCM mode," Intel Corporation, Tech. Rep., 2011.

0018-9448 (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/TIT.2019.2925613, IEEE Transactions on Information Theory

Ashwin Jha received his B.E. degree in computer engineering from the Delhi College of Engineering, Delhi, India in 2012, and M.Tech. degree in computer science from the Indian Statistical Institute, Kolkata, India in 2015. Currently, he is a Ph.D. student at the Indian Statistical Institute, Kolkata, India. His research interests include cryptology and pseudorandomness.

Cuauhtemoc Mancillas-López received his B.E. degree in electronic and communications engineering from ESIME-IPN, Mexico, in 2004, and M.Sc. and Ph.D. degree in computer science from CINVESTAV-IPN, Mexico, in 2007 and 2013 respectively. He did a post-doc in the Hubert Curien Laboratory at the University of Lyon at Saint Etienne, France. Currently, he is a researcher in the Department of Computer Science, CINVESTAV-IPN, Mexico City. His current research interests include design and analysis of provably secure symmetric encryption schemes, efficient software/hardware implementations of cryptographic primitives, and computational arithmetic.

Mridul Nandi received his B.Stat., M.Stat., and Ph.D. degree in computer science from the Indian Statistical Institute, Kolkata, India in 1999, 2001, and 2005, respectively. Currently he is an associate professor at the Applied Statistics Unit, Indian Statistical Institute, Kolkata, India. His research interests include design and analysis of symmetric-key schemes; efficient software/hardware implementations of cryptographic algorithms; functional encryption schemes and signature schemes.

Sourav Sen Gupta received his B.E. degree in electronics and telecommunication engineering from Jadavpur University, India, in 2006, M.Math. degree in pure mathematics from University of Waterloo, Canada, in 2008, and Ph.D. degree in computer science from Indian Statistical Institute in 2014.

From 2014 to 2017, he was a Lecturer cum Post-Doctoral Fellow with the R C Bose Centre for Cryptology and Security, Indian Statistical Institute, Kolkata. Since 2018, he has been a Lecturer with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. His research interests include cryptanalysis, cryptographic design, cryptocurrencies, blockchain technology, and applications of machine learning in cybersecurity.

0018-9448 (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.

arXiv:1907.00625v1 [cs.NE] 1 Jul 2019

On-chip learning in a conventional silicon MOSFET based Analog Hardware Neural Network

Nilabjo Dey*[,1], Janak Sharda*[,1], Utkarsh Saxena[1],
Divya Kaushik[1], Utkarsh Singh[2] and Debanjan Bhowmik[1]
*These authors contributed equally to the work.
[1]Department of Electrical Engineering,
Indian Institute of Technology Delhi, New Delhi 110016, India,
[2]Department of Electronics and Communications Engg.,
Delhi Technological University, New Delhi 110042, India
Email: debanjan@ee.iitd.ac.in

Abstract

On-chip learning in a crossbar array based analog hardware Neural Network (NN) has been shown to have major advantages in terms of speed and energy compared to training NN on a traditional computer. However analog hardware NN proposals and implementations thus far have mostly involved Non Volatile Memory (NVM) devices like Resistive Random Access Memory (RRAM), Phase Change Memory (PCM), spintronic devices or floating gate transistors as synapses. Fabricating systems based on RRAM, PCM or spintronic devices need in-house laboratory facilities and cannot be done through merchant foundries, unlike conventional silicon based CMOS chips. Floating gate transistors need large voltage pulses for weight update, making on-chip learning in such systems energy inefficient. This paper proposes and implements through SPICE simulations on-chip learning in analog hardware NN using only conventional silicon based MOSFETs (without any floating gate) as synapses. We first model the synaptic characteristic of our single transistor synapse using SPICE circuit simulator and benchmark it against experimentally obtained current-voltage characteristics of a transistor. Next we design a Fully Connected Neural Network (FCNN) crossbar array using such transistor synapses. We also design analog peripheral circuits for neuron and synaptic weight update calculation, needed for on-chip learning, again using conventional transistors. Simulating the entire system on SPICE circuit simulator, we obtain high training and test accuracy on the standard Fisher's Iris dataset, widely used in machine learning. We also account for device variability and noise in the circuit, and show that our circuit still trains on the given dataset. We also compare the speed and energy performance of our transistor based implementation of analog hardware NN with some previous implementations of NN with NVM devices and show comparable performance with respect to on-chip learning. Easy method of fabrication makes hardware NN using our proposed conventional silicon MOSFET really attractive for future implementations.

1 Introduction

Neural network (NN) algorithms are being widely used by the machine learning and data sciences community currently to solve several data classification and regression tasks [1]. However implementing NN on a traditional computer built on von Neumann architecture (memory and computing physically separated) involves continuous transfer of information between the memory and computing units. This von Neumann bottleneck leads to lower performance in terms of speed and energy consumption [2, 3, 4, 5, 6, 7]. Hence researchers have come up with specialized hardware NN implementations to get rid of the von Neumann bottleneck [8, 9, 10, 11]. Among these implementations, analog hardware NN uses a crossbar array of synaptic devices to perform computing at the location of the data itself [5, 6, 12]. The fact that such crossbar array enables execution of Vector Matrix Multiplication (VMM), inherent in a FCNN algorithm, in a parallel fashion makes it suitable both for forward inference [5] or on-chip learning (training in hardware). In fact, on-chip learning in such crossbar array has been considered to be faster and more energy efficient than conventional training of NN on GPU [12].

A synaptic device in a crossbar array based analog hardware NN must have several conductance states that can be controlled electrically to store and update the weight values of the NN. Several Non Volatile Memory (NVM) devices like floating gate transistors, chalcogenide based Phase Change Memory (PCM) devices, oxide based Resistive Random Access Memory (RRAM) devices and spintronic devices have been proposed and used as synaptic devices in previous implementations of analog hardware NN [5, 6, 7, 13, 14, 15, 16, 17, 18, 19]. However, these NVM devices have several issues associated with them with respect to achieving on-chip learning in crossbar arrays made of them. Floating gate transistor synapses need high voltage pulses for weight update and have low endurance [19, 20, 21, 22, 23, 24]. Memristive oxide based Resistive Random Access Memory (RRAM) devices and Phase Change Memory (PCM) devices [16, 17, 18] exhibit an asymmetric/unipolar and non-linear dependence between conductance (and weight) update and programming pulses, which affects the accuracy during on-chip learning of crossbar arrays that use such devices [5, 25, 26, 27]. Moreover, for fabricating RRAM, PCM or spintronic device based hardware NN systems [5, 7, 28], dedicated in-house fabrication facilities are needed since they involve novel materials. The system cannot be designed in-house and then fabricated elsewhere e.g. in commercial merchant foundries, unlike silicon based conventional CMOS circuits and systems.

Instead if conventional Metal Oxide Semiconductor Field Effect Transistors (MOSFETs) with silicon (Si) as semiconducting material and SiO_2 as gate oxide could be used as synaptic elements and analog values of weight could be stored in them, unlike what's done in an SRAM cell where digital bits are stored, then fabrication of analog hardware NN will be much easier. In this paper, we propose such a conventional MOSFET as a three terminal synaptic element in analog hardware NN. In Section 2, we show through SPICE simulations at 65 nm technology node on Cadence Virtuoso circuit simulator that the conductance between drain and source of the transistor (first and second terminal) can be controlled between several analog values, which represent the weight values, by changing the voltage applied at the gate of the transistor (third terminal) (Fig. 1). We benchmark our data against experimentally measured data on a n-MOSFET present inside a commercially available chip (Fig. 2) Gate voltage can be changed by applying current pulses at the gate and charging/ discharging the gate oxide, as we show in our simulations (Fig. 3). Conductance (weight) vs programming gate current pulse plot is found to be fairly linear and symmetric for positive and negative weight update for our proposed transistor synapse (Fig. 3), unlike PCM and RRAM based synaptic devices in which non-linearity and asymmetry/ unidirectionality in

conductance response degrades the overall neural network performance [5, 25, 26, 27].

In Section 3 we design a Fully Connected Neural Network (FCNN) circuit in crossbar topology using such synaptic transistors at 65 nm technology node that carries out the input Vector- weight Matrix Multiplication (VMM), characteristic of FCNN (Fig. 4). We design, using transistor and transistor based op-amps, analog neuron circuit (Figure 5) and Stochastic Gradient Descent (SGD) algorithm [1] based synaptic weight update calculation circuit (Figure 6). This weight update circuit sends current pulses to the gates of the synaptic transistors. These gate current pulses update the weight values of the FCNN (Fig. 3, 4).

In Section 4, using SPICE simulations of the whole system on Cadence Virtuso circuit simulator, we demonstrate on-chip learning in our proposed hardware (Fig. 7) on the Fisher's Iris dataset- a popular machine learning dataset (Fig. 9) [29]. In Section 5, we next compare the speed and energy performance of our transistor synapse based analog NN against analog implementations of the same through previously proposed NVM devices with respect to on-chip learning on the exact same dataset [17, 25, 30, 49, 31]. Speed and energy consumption are almost equal for our proposed transistor synapse and spintronic (domain wall based) synapse (Table I). Compared to RRAM synapse, speed for transistor synapse is much higher. Also energy is several orders lower (Table I), again because of the asymmetric nature of conductance response of RRAM synapse as opposed to the synaptic nature of the same in transistor synapse. In Section 6, we argue that our proposed transistor synapse based FCNN circuit trains itself even in the presence of synaptic device variability and noise in input voltages.

Thus we show that Si MOSFET synapse can be considered as an attractive candidate for implementation of hardware analog NN. To the best of our knowledge, this is the first proposal and demonstration through simulations of on-chip learning on analog hardware NN using only conventional Si MOSFETs as synapses. It is to be noted that [19, 20, 21, 22] propose synaptic behaviour of MOSFET by using a floating gate. However weight modulation in floating gate synapse is much slower because electrons have to be injected inside the gate through a tunneling mechanism for a change of weight. Also, the voltages needed for such weight update are very large [23].Number of times charge can be injected into/ ejected out of the floating gate is also limited, leading to low endurance [24]. Since weight has to be frequently updated for on-chip learning, floating gate transistor hence doesn't make for a good synapse. On the other hand, our synaptic MOSFET is conventional, doesn't have a floating gate and hence doesn't suffer from those disadvantages.

It is also to be noted that earlier reports of conventional silicon transistor based synapse use multiple transistors to store each bit of the weight value stored in the synapse [32, 33]. On the other hand, analog values of weight are stored in a single transistor in our proposed synapse as different conductance states.

Also, in this paper, we demonstrate the capability of our transistor synapse in a non-spiking network, trained "on-chip" through the much developed Stochastic Gradient Descent (SGD) algorithm benchmarked on several machine learning datasets, as opposed to spiking network which mostly uses Spike Time Dependent Plasticity (STDP) enabled training algorithm [25, 34, 35, 36, 38, 39]. Convergence properties and highly accurate training results have not been demonstrated on large datasets in STDP enabled spiking network algorithms to the extent they have been in SGD based non-spiking network algorithms [36, 37].

2 Characterization of a single conventional MOSFET as a synaptic device

Fig. 1(a) shows schematic of a Si-SiO_2 based n-MOSFET we propose as synapse in this paper.

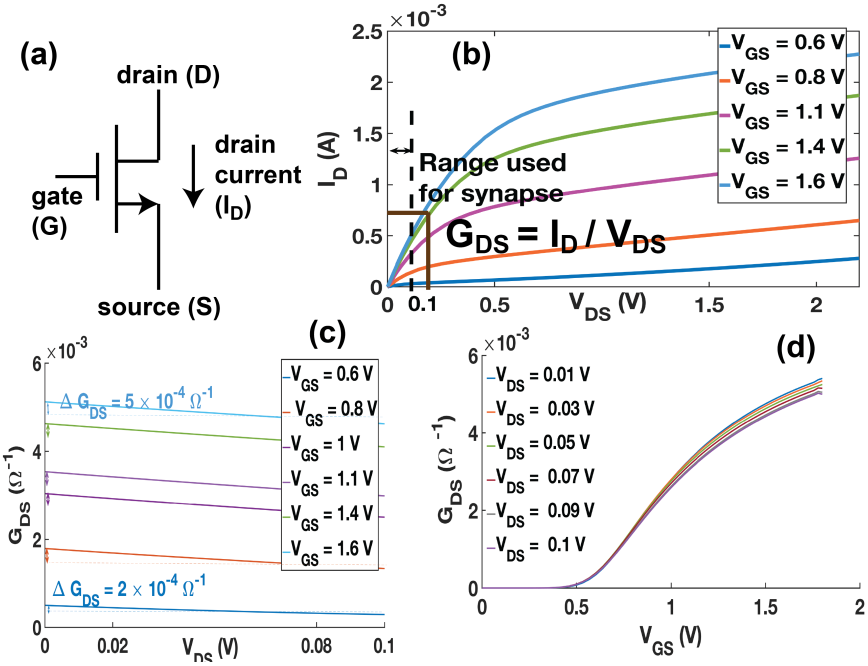


Figure 1: (a) Schematic of proposed n-MOSFET synapse. Following sub-plots are obtained from simulation of the MOSFET at 65 nm node on SPICE circuit simulator. (b) Drain current (I_D) as a function of drain to source voltage (V_{DS} for different applied gate to source voltages (V_{GS}) (c) Conductance between drain (D) and source (S) (G_{DS}) plotted as a function of V_{DS} for different values of V_{GS}. (d) (G_{DS}) plotted as a function of V_{GS} for different values of V_{DS}.

We simulate it at the 65 nm technology node through SPICE simulations on Cadence Virtuoso using the United Microelectronics Corporation (UMC) library. Drain current (I_D) vs drain to source voltage (V_{DS}) characteristic is linear for a certain range of V_{DS} (0 - 0.1 V in this case) for gate to source voltage V_{GS} in between 0.6 V and 1.6 V (Fig. 1(b)). This behavior is expected for conventional MOSFET [40] and matches qualitatively with I_D- V_{DS} characteristic we experimentally measure on a single n-MOSFET present inside the commercially available CD4007 inverter chip and accessible through package terminals (Fig. 2(a)). We operate V_{DS} between 0 and 0.1 V, and V_{GS} between 0.6 and 1.6 V for functioning as synapse throughout this paper.

For any combination of V_{GS} and V_{DS}, ratio of I_D to V_{DS} determines drain to source conductance (G_{DS}), which is a function of both V_{GS} and V_{DS} ($G_{DS}(V_{GS}, V_{DS})$) (Fig. 1(b)). We observe that for a fixed V_{GS}, change in G_{DS} (ΔG_{DS}) is in the order of $10^{-4}\ \Omega^{-1}$ when V_{DS} varies in the full range we have selected (0 - 0.1 V) (Fig. 1(c)). However, for a fixed V_{DS} when V_{GS} varies in full range (0.6 - 1.6 V), change in G_{DS} is $5{\times}10^{-3}\Omega^{-1}$, which is one order higher than change in G_{DS} due to full sweep of V_{DS}. Thus G_{DS} can just be approximated as a function of V_{GS} and not V_{DS} (and in extension I_{DS}) in the selected range of operation.

Hence we can write

$$I_D = G_{DS}(V_{GS}, V_{DS})V_{DS} \approx G_{DS}(V_{GS})V_{DS} \quad (1)$$

Experimentally measured data on transistor

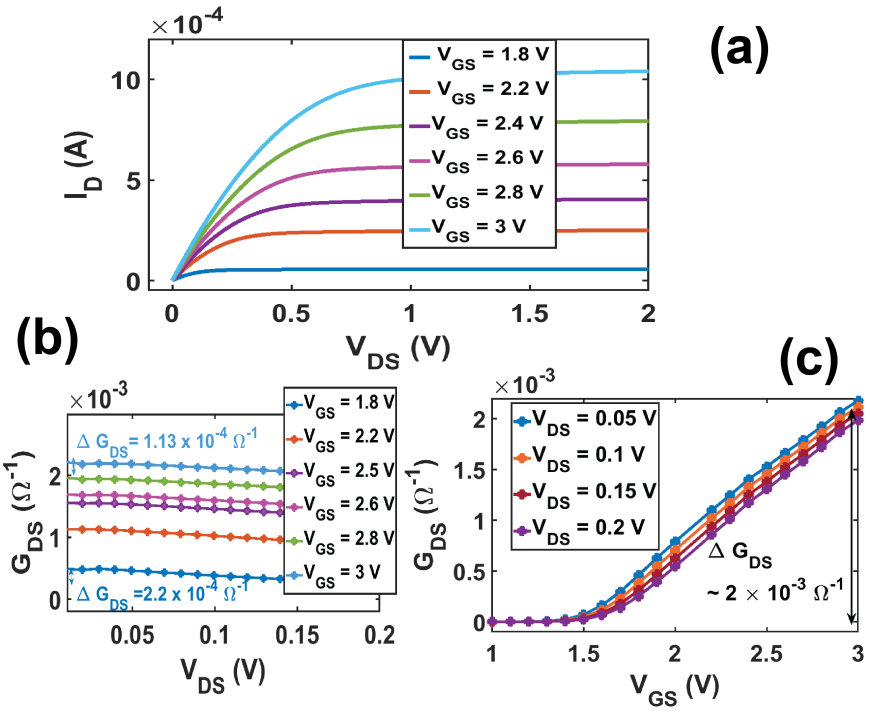


Figure 2: Experimental data obtained from measurement on single n-MOSFET in commercially available CD4007 chip (a) Drain current (I_D) as a function of drain to source voltage (V_{DS} for different applied gate to source voltages (V_{GS}) (b) Conductance between drain (D) and source (S) (G_{DS}) plotted as a function of V_{DS} for different values of V_{GS}. (c) (G_{DS}) plotted as a function of V_{GS} for different values of V_{DS}.

in CD4007 chip qualitatively matches with this observation (Fig. 2(b),(c)). For a quantitative match, the specifications of the simulated and experimentally measured transistors need to match which has not been the case in this work. When our simulated transistor is used as synapse in M input nodes $\times$ N output nodes crossbar array based analog hardware FCNN as shown in Fig. 4(a), input vector- weight matrix multiplication, or VMM operation, takes place as a part of the feedforward computation both for the training phase and testing/ inference phase [5]. During this process, the input vector $\{x_1, x_2, x_3...x_m..x_M\}$ corresponding to a training sample acts as drain voltages on the transistor synapses as shown in Fig. 4(a). The sources of all the transistor synapses are maintained at 0 V using op-amps at the input stage of the neuron circuits we design (Fig. 5). Thus for a transistor synapse connecting input node m with output node n, its V_{DS} is proportional to input x_m. If its conductance G_{DS} represents its weight $w_{n,m}$ then from equation (1) its drain current I_D turns out to be proportional to $w_{n,m}x_m$. Even when V_{DS} and hence I_D changes during the VMM operation in both training and testing phase of the hardware, as long as V_{DS} stays in the chosen range of 0 and 0.1 V this relation holds true. In any case, since we show training through SPICE simulations of the entire hardware in this paper,

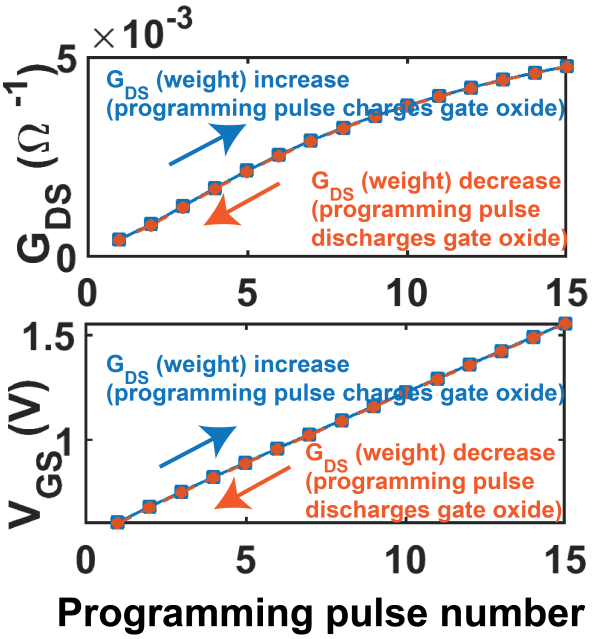


Figure 3: Conductance G_{DS} (and hence weight)and corresponding gate voltage (V_{GS}) increase linearly due to programming current pulses which charge the capacitive gate oxide of the MOSFET synapse (blue plot). Conductance and gate voltage decrease linearly due to programming current pulses of opposite polarity which discharge the gate oxide (orange plot)

any effect of higher order term of drain voltage V_{DS} on the current I_D is taken into account in our analysis and final accuracy results.

Conductance (G_{DS}) and hence synaptic weight update is carried out by applying current pulses at the gate of the MOSFET synapse and charging/ discharging the gate oxide. Fig. 3 shows that when pulses of the same current magnitude (132 nA) and same duration (1 ns) with polarity such that gate oxide charges up, gate voltage V_{GS} and hence G_{DS} increases linearly with pulse number (blue plot). Similarly starting from the highest G_{DS}, programming pulses of opposite polarity but of same magnitude and duration discharge the oxide and G_{DS} decreases (orange plot). This linear and symmertrical/bipolar nature of weight update with programming pulse number is not the case in RRAM and PCM based synaptic devices, making implementation of on-chip learning in crossbar arrays made of such synapses quite challenging [5, 25, 26]. Even when a pair of such RRAM or PCM devices is used as a single synapse as has been done before, frequent RESET pulses, which are of long time duration and consume a lot of energy, are still needed to carry out the weight update scheme [18, 25].

The weight update process for our proposed transistor synapse is repeated for every training sample in each epoch to obtain high training accuracy after a certain number of epochs and thus achieve on-chip learning, as we show next.

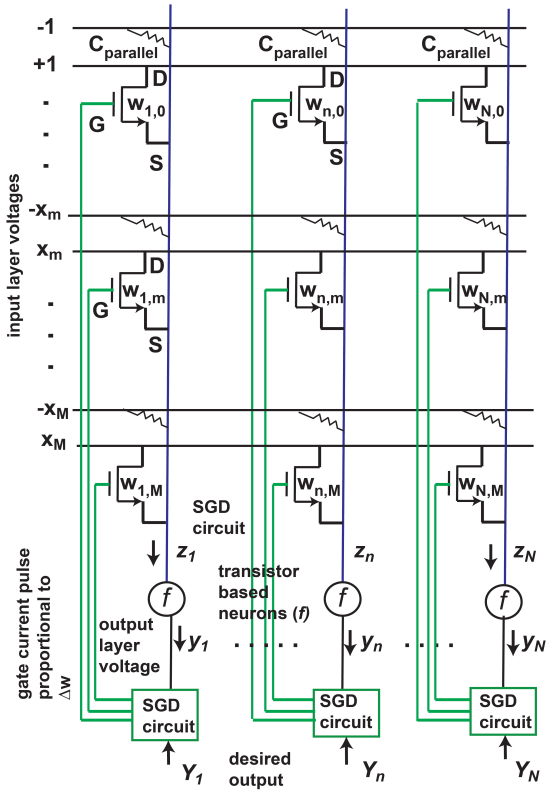


Figure 4: Schematic of transistor synapse based cross-bar architecture for weight matrix-input vector multiplication, followed by transistor based neuron circuit for activation function (f) (shown in Figure 5) and then feedback circuit that evaluates change in weight needed based on output of neurons and desired output using Stochastic Gradient Descent (SGD) method (shown in Figure 6). The feedback circuit then applies current pulses at the gates of the transistor synapses such that the gate voltage, and hence the weight stored by the transistor synapse, changes by the desired amount

3 Design of crossbar Fully Connected Neural Network (FCNN) circuit and feedback circuit to train it

Analog crossbar array of our proposed synaptic transistors is designed in Cadence Virtuoso SPICE circuit simulator next to implement a Fully Connected Neural Network (Fig. 4) [1, 41]. If the input layer has M nodes and output layer has N nodes then for an input vector $\{x_1, x_2, x_3...x_m..x_M\}$ at the input layer the output vector at the output layer is given by $\{y_1, y_2, y_3...y_n..y_N\}$ where:

$$y_n = f(z_n) = \frac{2}{1 + e^{-\lambda z_n}} - 1 \tag{2}$$

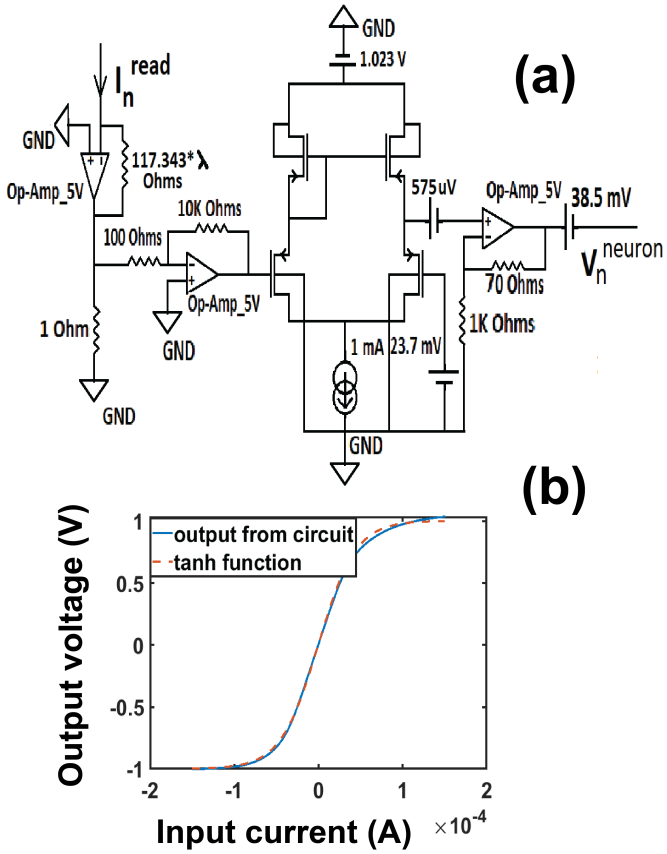


Figure 5: (a) Neuron circuit present at each output node of the crossbar circuit in Figure 4 shown in details here. (b) Output voltage of neuron circuit as a function of input current , obtained from SPICE simulation is shown with solid line (blue). Ideal tanh function is plotted with dashed line (red)

and

$$z_n = w_{n,1}x_1 + w_{n,2}x_2 + ... + w_{n,m}x_m + ..w_{n,M}x_M + w_{n,0} = (\Sigma_{m=1}^{m=M} w_{n,m}x_m) + w_{n,0} \quad (3)$$

where $w_{n,m}$ is the weight of the synapse connecting input node m with output node n, f is the tanh activation function and λ is a parameter in the function. As discussed in the previous section, to implement the weight matrix-input vector multiplication of equation (3) in hardware, we map the weight to the conductance between drain and source of MOSFET synapse (G_{DS}). Minimum value of weight corresponds to minimum conductance value and maximum value of weight corresponds to maximum conductance value within the chosen gate voltage range where conductance is almost linearly proportional to gate voltage (Fig. 2(c)). Voltage proportional to input at node m (x_m) is applied at the drain (V_{DS}) of transistor synapse connecting input node m with output node n (Fig. 4), in the form of a 1ns long pulse corresponding

to one training sample. V_{DS} corresponding to maximum value of x_m is 0.1 V. G_{DS} varies little with V_{DS} when V_{DS} is below 0.1 V as seen in our simulations (Fig. 2(b)).

Since weight takes both positive and negative values while conductance is only positive, a resistance is added in parallel to a transistor synapse and voltage proportional to negative of x_m is applied to it (Fig. 4). Thus current proportional to $w_{n,m}x_m$ flows between drain and source of the transistor. Currents of all transistors connected to output node n add up following Kirchoff's Current law generating current proportional to z_n in equation (3).

Current proportional to z_n next enters a transistor based analog neuron circuit that executes tanh activation function (f) of equation (2) [42, 43]. Thus output voltage of neuron circuit corresponds to y_n at output node n (Fig. 5). We design this neuron circuit using a set of pre-amplifiers composed of transistor based op-amps and a transistor based differential amplifier circuit, again on Cadence Virtuoso simulator (Fig. 5(a)). Output voltage vs input current plot (Fig. 5(b)) for the neuron circuit shows desired tanh behavior of equation (1). Hyperparameter (λ) in the function (equation 2) can be adjusted by changing a resistance in the neuron circuit as shown in Fig. 5(a)).

Next this voltage is applied to a transistor based feedback circuit, also analog just like the neuron circuit. This circuit present at every output node updates the weights and trains the network in hardware, or rather accomplishes on-chip learning via Stochastic Gradient Descent (SGD) calculation as discussed below (Fig. 6).

For almost every vector in the input set for which the training is done (training set) $\{x_1, x_2, x_3...x_m..x_M\}$ the output $\{y_1, y_2, y_3...y_n..y_N\}$ must match the desired output: $\{Y_1, Y_2, Y_3...Y_n..Y_N\}$, corresponding to the known label/ class that input vector belongs to. In order to do that, the weight matrix need to be updated adopting the Stochastic Gradient Descent (SGD) method through several iterations over the training set.

The error generated at output node n is given by:

$$\epsilon_n = \frac{1}{2}(Y_n - y_n)^2 \tag{4}$$

Weight of the synapse connecting output node n with input node m ($w_{n,m}$) is updated between iteration i and $i+1$ as follows:

$$\begin{aligned} w_{n,m}^{i+1} &= w_{n,m}^{i} - \Delta w_{n,m} \\ &= w_{n,m}^{i} - \eta \frac{\partial \epsilon_n}{\partial w_{n,m}} \\ &= w_{n,m}^{i} - \frac{\eta\lambda}{2}(Y_n - y_n)(1 - y_n^2)x_m \end{aligned} \tag{5}$$

and weight of the bias synapse for output node n is updated as follows:

$$\begin{aligned} w_{n,0}^{i+1} &= w_{n,0}^{i} - \Delta w_{n,0} \\ &= w_{n,0}^{i} - \eta \frac{\partial \epsilon_n}{\partial w_{n,0}} \\ &= w_{n,0}^{i} - \frac{\eta\lambda}{2}(Y_n - y_n)(1 - y_n^2) \end{aligned} \tag{6}$$

where η is the learning rate and each iteration corresponds to each training sample inside each epoch. Learning rate and the hyperparameter λ can be adjusted in hardware, by changing the gain of an opamp, using a variable resistor in the "amplifier" block of the SGD calculation circuit (Fig. 6(a)). The SGD circuit has been designed by us on Cadence Virtuoso using transistor and op-amp (made of transistors) based subtractor and multiplier blocks as shown in Fig. 6(b,c,d). The SGD circuit computes the weight update ($\Delta w_{n,m}$) [49]. Building a subtractor block from op-amp is a standard process in analog electronics [44]. The multiplication operation is carried out with a single transistor making use of the fact that I_{DS} is proportional to V_{DS} times V_{GS}[37].

Since weight of each transistor synapse is proportional to the applied gate to source voltage (V_{GS}) (Fig. 1(b)), in order to update the weight at each synapse, from equation (4) and (5) the gate voltage has to be updated as follows:

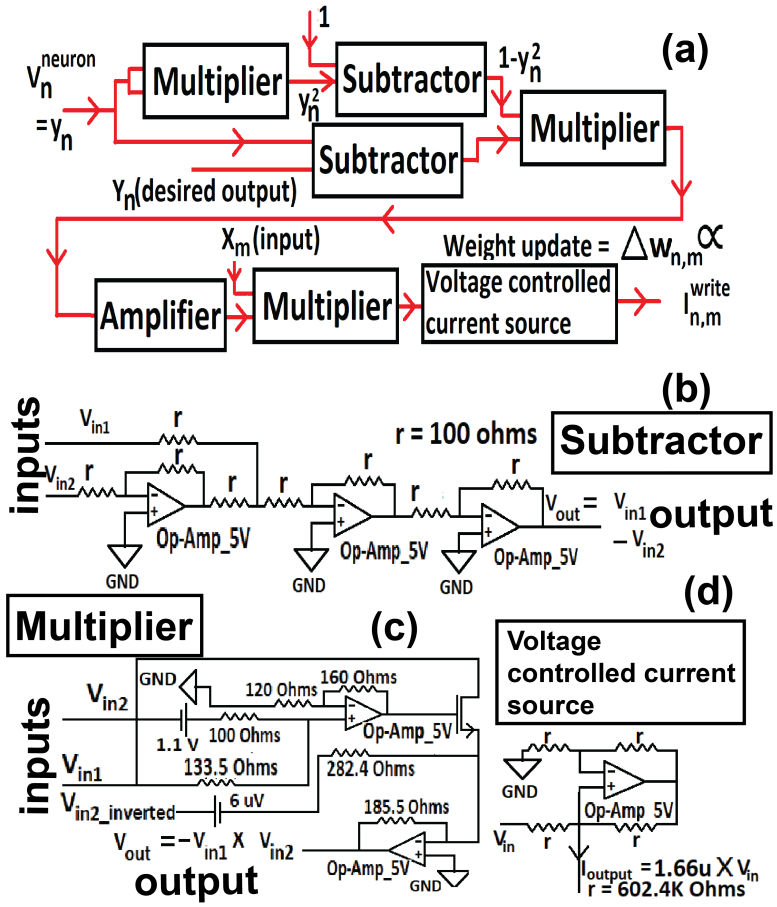


Figure 6: (a) Analog peripheral feedback circuit we designed, using conventional silicon transistors and op-amps made of them, on circuit simulator for weight update following SGD algorithm (b) Design of subtractor block (c) Design of multiplier block (d) Design of voltage controlled current source circuit.

$$V_{GS,n,m}^{i+1} = V_{GS,n,m}^{i} - \Delta V_{GS,n,m};$$
$$V_{GS,n,0}^{i+1} = V_{GS,n,0}^{i} - \Delta V_{GS,n,0} \quad (7)$$

The weight update is calculated ($\Delta w_{n,m}$) by the SGD calculation circuit and generated in the form of a 1 ns long voltage pulse since input voltage pulse is 1ns long, corresponding to a training sample. A voltage controlled current source, working based on the principle of Howland current pump [45], is designed at the output stage. It converts the voltage pulse to 1 ns long programming current pulse. Magnitude of current is proportional to $\Delta w_{n,m}$ (Fig. 6(a),(d)). The current pulse is applied at the gate of the transistor such that change in gate voltage is proportional to the integral of gate current over time, as shown in the previous section, and is equal to the weight update of equation (6). Such weight

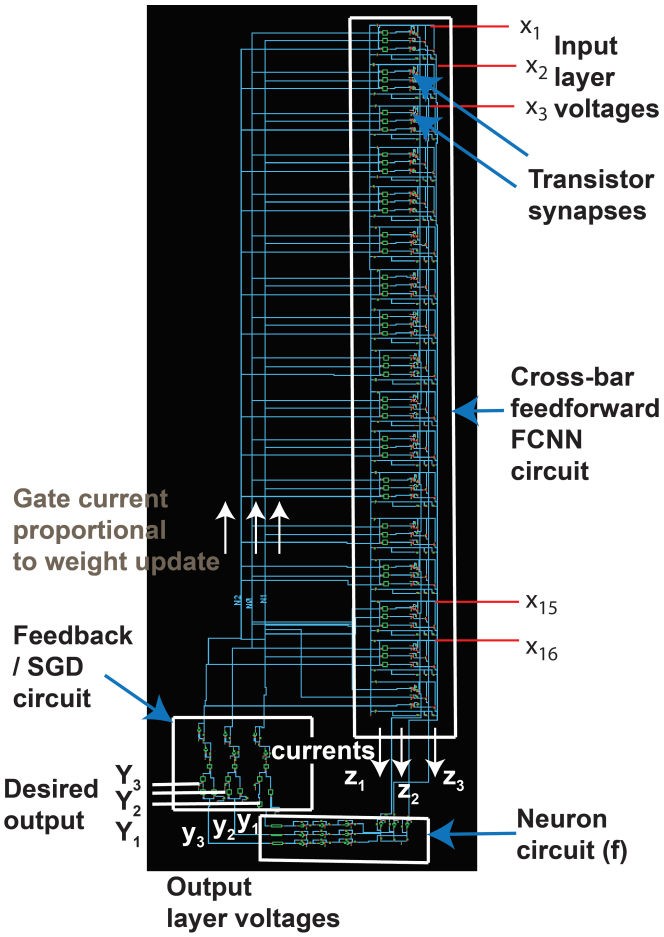


Figure 7: Schematic of crossbar circuit of Fig. 4, neuron circuit of Fig. 5 at each output node of crossbar circuit and feedback circuit (SGD algorithm based) of Fig. 6 at output node of crossbar circuit all wired together. The schematic is designed and simulated on Cadence Virtuoso (SPICE) circuit simulator for on-chip learning on Fisher's Iris dataset.

update carried out over all transistor synapses over all training samples through several repetitions/ epochs results in on-chip learning of the designed FCNN.

4 Accuracy, energy and speed analysis for on chip learning on Fisher's Iris dataset

We next do SPICE simulations of the transistor synapse based crossbar FCNN circuit (Fig. 4), neuron circuit (Fig. 5) and SGD based feed-

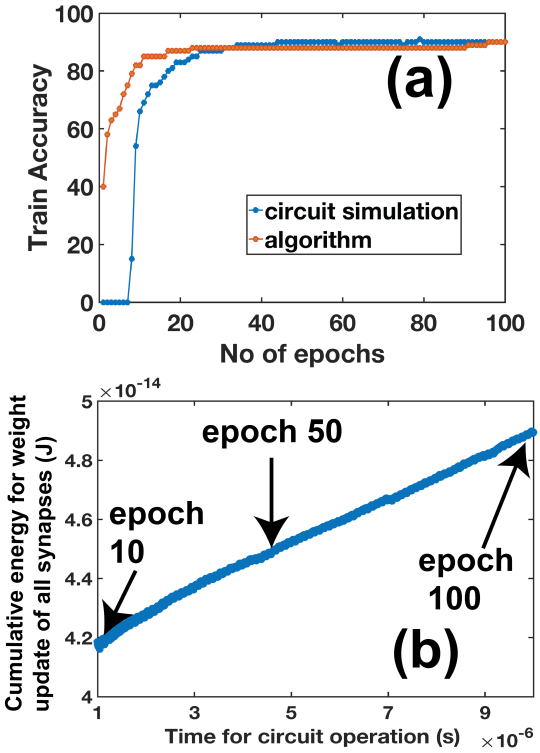


Figure 8: (a) Accuracy as a function of epochs on 100 train samples of Fisher's Iris dataset, obtained from SPICE simulations of circuit in Figure 7 (orange plot) and execution of the same algorithm on a conventional computer (software neural network). (b) Corresponding training energy (cumulative) in the designed circuit vs epoch is plotted. It includes energy consumption for weight update of all the synapses.

back/ weight update circuit (Fig. 6) we design, wired all together as shown in Fig. 4. The corresponding schematic of Fig. 4 on the SPICE simulator (Cadence Virtuoso) is shown in Fig. 7. We carry out on-chip learning for the circuit, following the method described in the previous section, on the Fisher's Iris dataset- a standard dataset in the machine learning community [29].

There are 16 input nodes in our FCNN circuit (Fig. 7) corresponding to 16 inputs: 4 features of flowers passed through 4 sensors each. [48]: $f(x) = x$, $f(x) = 1 - x$, $f(x) = 1 - 2|x - 0.5|$, $f(x) = 2|x - 0.5|$. There are 3 output nodes corresponding to 3 possible classes of flowers. For flower of type A, desired output $Y_1, Y_2, Y_3 = 1, -1, -1$. For flower of type B, desired output $Y_1, Y_2, Y_3 = -1, 1, -1$. For flower of type C, desired output $Y_1, Y_2, Y_3 = -1, -1, 1$. The available dataset has total 150 samples, almost equally distributed among types A,B and C. 100 samples are used here for training in each epoch and 50 separate samples are used for final testing to determine test accuracy.

Each sample is trained for 1 ns. Our transistor synapses can be operated at higher speed but magnitude of gate currents for weight updates and hence total energy for training will also go up. So we choose this speed of 1 ns per training

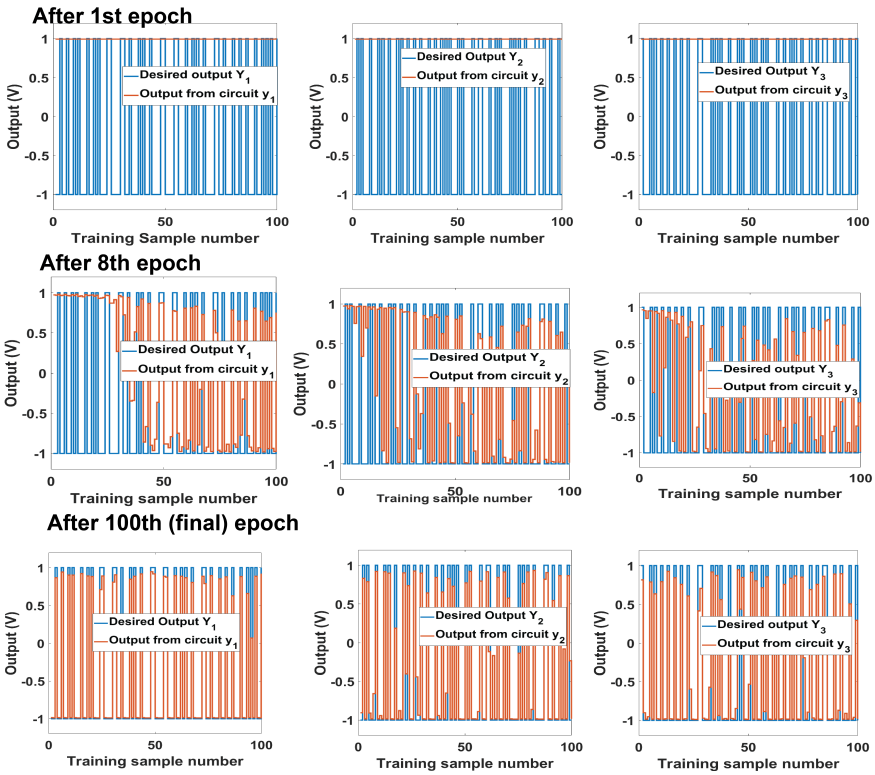


Figure 9: Output voltage of neuron circuit at each node of the circuit in Figure 6 , obtained from SPICE simulation, is plotted with respect to every training sample in an epoch. Results for 2nd epoch (a), 10th epoch (b) and 100th epoch (c) are shown.

sample in an epoch.Training accuracy vs epoch plot is shown in Fig. 8 (a). Accuracy number is obtained by comparing the voltage waveforms at the output nodes, i.e. at the output stage of the neuron circuits (y_1, y_2, y_3), with the desired waveforms (Y_1, Y_2, Y_3), from the SPICE simulation of our full neural network circuit shown in Fig. 7 . For every sample, if the output at each of the three nodes is within 40 percent of the desired output then we consider it a success. For about the first five epochs the accuracy is 0. This is because all three output nodes y_1, y_2, y_3 are at 1,1,1 V for all samples. Hence the output is wrong for all the samples. Output for 2nd epoch for example is shown in Fig. 9(a). Around 10th epoch the output nodes start giving correct output and hence accuracy suddenly increases (Fig. 8(a)). Output waveform for 10th epoch is shown in Fig. 9(b). Beyond the 10th epoch, outputs at all nodes gradually start following the desired outputs. Thus accuracy increases with epoch as shown in Fig. 8(a). Output waveform for 100th epoch shows that all output nodes give the same output as desired for 90 samples. Thus after 100 epochs, or 100 x 100 iterations or 10 μs, the accuracy(both train and test) (on 100 train samples) is 90 percent which is very similar to algorithm implemented in python code. Thus we have been able to achieve successful on-chip learning of our transistor synapse based analog FCNN circuit in SPICE simulations.

From Fig. 8, by 50th epoch or 5 μs,

Table 1: Comparison of performance for on-chip learning of our transistor synapse based FCNN circuit with implementation of the same FCNN using other synaptic devices on the same dataset (Fisher's Iris)

Type of synapse	Time per sample per epoch	Total time for learning	Total energy consumed in synapses while learning
Transistor (this work)	1 ns	10 μs	50 fJ
Domain wall based spintronic device	3 ns	30 μs	9 fJ
Oxide based RRAM device	between 200 ns (each pulse for conductance increase) and 6 μs ("Reset" pulse)	variable	1 μJ

train(and test) accuracy reaches 90 percent(b) and net energy consumed in the synapses is as low as $\approx$ 45 fJ. Energy consumed in 100 epochs is $\approx$ 50 fJ. It is to be noted that energy consumed in analog peripheral circuits is ignored in this calculation. Those circuits can be further optimized to enable ultra low energy on-chip learning on our proposed transistor synapses based analog hardware NN.

5 Comparision of performance of proposed MOSFET synapse with other kinds of synaptic devices

Next we compare the energy and speed performance of on-chip learning on our proposed MOSFET synapse based FCNNN with that on FCNN designed using some other kinds of synaptic devices that have been proposed and implemented elsewhere - spin orbit torque driven domain wall based synapses (spintronic synapses) [7, 49, 50] and memristive oxide based RRAM synapses [51, 52]. For fair comparison between the different synaptic devices, neural networks with the same architecture and number of nodes needs to be designed with different types of devices as synapses and they have to be trained on the same dataset using the same algorithm. Hence we simulate the same 16 input node x 3 output node FCNN circuit of Fig. 7 with transistor synapses, domain wall synapses and RRAM based synapses. We use the same rule for weight update (Fig. 6) to train these three FCNN circuits. Time for training each sample in each epoch (each iteration) during on-chip learning, total time for the learning and energy consumed in the process are listed in Table 1 for the three FCNN circuits. The energy listed includes only energy consumed in the synapses for weight update during on-chip learning.

Our domain wall synapse model is micromagnetic in nature and is calibrated against experiments of current driven domain wall motion in Pt/Co/MgO devices [53, 54]. More details on our method for on-chip learning in domain wall synapse based FCNN circuit can be found in [49]. Time for learning is comparable between transistor synapse FCNN and domain wall synapse FCNN circuit. Energy consumed is approximately 5 times higher for transistor synapses compared to domain wall synapses (Table I). However our transistor synapses are much easier to fabricate because they involve conventional silicon MOSFET, which is not the case with domain wall synapses which needs magnetic materials sputtered under specific conditions so that they exhibit perpendicular magnetic anisotropy and high Dzyoloshinski Moriya interaction [28, 53, 54].

Speed and energy performance of our designed transistor synapses based NN is also compared with similar NN designed using HfO_x based RRAM synapse (Verilog model from [51] used in our simulation) with respect to on-chip learning on the same Fisher's Iris dataset in Table I. Though two RRAM devices are used per synapse in our simulation to address the asymmetry issue in conductance response [18, 25, 52], several long duration (6 μs) "Reset" pulses are still needed for successful weight update [25]. Hence a lot of energy is consumed for RRAM based NN compared to our transistor synapse based NN which doesn't have that issue, as explained earlier [25]. Also, since "Reset" pulse may or may not be needed for a particular synapse while training the NN on each sample in each epoch, the net time for training varies in the case of RRAM based NN in our weight update scheme. But time for training RRAM synapse based NN is certainly longer than our proposed transistor synapse based NN since even one "Reset" pulse is about a μs long while any pulse to increase or decrease weight in our transistor synapse is just 1 ns long, owing to its linear, bipolar synaptic characteristic (Fig. 3). Also, fabricating a RRAM based device requires in-house facilities and cannot be fabricated as easily as our proposed transistor synapse.

.

6 Device variability, noise in circuit voltages and retention of weights

Since there can be device level variations among the different transistor synapses, their weights can vary from the values calculated by the weight update circuit. Considering noise in the range of 0-10 percent for all the synaptic weights, train accuracy still turns out to be almost the same as ideal transistor synapses from our calculation (Fig. 10). This agrees with the observation in PCM synapse based FCNN, trained on-chip by similar gradient descent algorithm, that accuracy does not depend much on device stochasticity [25]. Similarly, we observe that noise in the training inputs through the drain voltages (Fig. 4) or error in weight update calculation up to 10 percent in the designed peripheral circuit does not affect the overall training accuracy much (Fig. 10).

Though linear synaptic characteristic and easy, well developed method of fabrication are the two main advantages of our proposed transistor synapse compared to RRAM, PCM or spintronic synapses, unlike those Non Volatile Memory (NVM) based synapses, our proposed transistor based synapses do not retain their weights for a very long time. Our SPICE simulations show that once on-chip learning is achieved in the designed transistor synapse based NN circuit (Fig. 4) and training inputs are stopped from being fed, the synaptic weights decay in about 1 ms. This is because the weight is proportional to the conductance, which is in turn proportional to the gate voltage. As the gate oxide capacitance of the synapse discharges, gate voltage drops and hence the correct weight value is lost with time. However, the retention time is 10^6 times higher than time for training each sample in an epoch (1 ns). For that retention factor, on-chip learning can still be achieved as argued in [46, 47] and seen in our simulations (Fig. 8). However this means that our proposed transistor synapse based NN is not suitable for applications where the training is done in a traditional computer and only testing/ inference is done in hardware (off-chip learning). On the other hand, in applications where training and testing both need to be done in hardware, our proposed system provides an easily fabricable platform for training in hardware (on-chip learning) as opposed to PCM or RRAM synapse based NN ([5]) and does not need high voltage for weight update like floating gate synapse based NN. For testing/inference after a certain time duration from training, the proposed solution is to store the final weights after training in a floating gate synapse based NN for testing purposes, where this needs to be done only once and thereby the endurance and energy

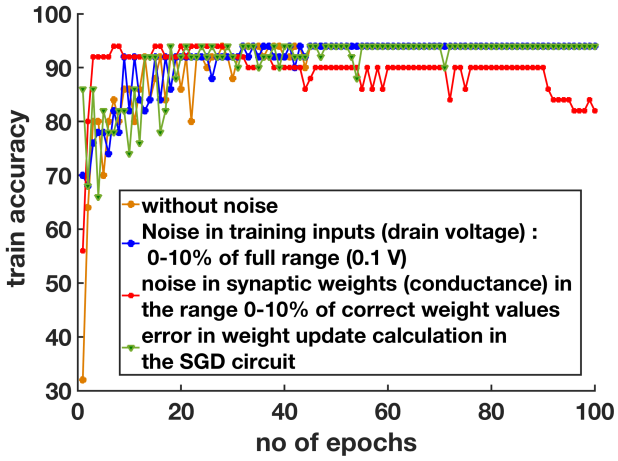


Figure 10: Training accuracy is plotted as function of epoch for the following four cases: without any noise in the circuit, in presence of noise in all training inputs through noise in drain voltages, in presence of noise in synaptic weights through device variation in conductances of transistor synapses and in presence of noise in SGD based weight update calculation in peripheral feedback circuit.

consumption issues in writing weight values on floating gate synapses do not become major bottleneck [23, 24].

7 Conclusion

Thus in this paper we have proposed a new functionality of conventional Si-SiO_2 based MOSFET as synaptic device in analog hardware NN. No floating gate is present in the synapse unlike previous proposals. Through SPICE simulations we demonstrate successful on-chip learning on Fisher's Iris dataset and compare its energy and speed performance with other implemenations of analog hardware FCNN with previously proposed spintronic and oxide based RRAM synapses. Easy means of fabrication through established silicon based merchant foundries and linear, bipolar synaptic characteristics make our MOSFET synapse a potential candidate for implementation of analog hardware NN in the near future.

Acknowledgment

The authors would like to thank Rajinder Singh Deol and Madhusudan Singh for help with the experimental measurement, and Apoorv Dankar, Anand Kumar Verma, Shouri Chatterjee and Ankesh Jain for help with the simulations.

References

[1] Y. LeCun *et al.*, Nature, vol. 521, pp 436-444, 2015.

[2] J. Misra and I. Saha, Neurocomputing, vol. 74, pp. 239-255, 2010.

[3] C. D. Schuman *et al.*, arxiv. 1705.06963v1.

[4] K. Boahen, Computing in Science & Engineering, vol. 19, no. 2, 2017.

[5] H. Tsai *et al.*, J. Phys. D: Appl. Phys. vol. 51, no. 283001, 2018.

[6] A. Sebastian *et al.*, J. Appl. Phys. vol. 124, no. 111101, 2018.

[7] A. Sengupta *et al.*, IEEE Transactions on Circuits and Systems- I, vol. 63, no. 12, 2016.

[8] P. A. Merolla *et al.*, Science vol. 345, no. 6197, pp. 668-697, 2014.

[9] S.B. Furber *et al.*, Proceedings of the IEEE, vol. 102, no. 5, pp. 652-665, 2014.

[10] M. Davies *et al.*, IEEE Micro., pp. 82-89, 2018.

[11] A. Diamond *et al.*, Frontiers in Neuroscience, vol. 9, pp. 491, 2016.

[12] G.W.Burr *et al.* in Proc. IEEE International Electron Devices Meeting (IEDM), pp. 29.5.129.5.4, 2014

[13] D. Ielmini and H. S. Philip Wong, Nature Electronics, vol. 1, no. 111101, pp. 333-343, 2018.

[14] M. Suri ed., Cognitive Systems Monographs, Springer, 2017.

[15] J. Grollier *et al.* Proceedings of the IEEE, vol. 104, no. 10, 2016.

[16] I. Boybat *et al.*, Nature Communications, vol. 9, no. 2514, 2018.

[17] Y. Li *et al.*, J. Phys. D: Appl. Phys. vol. 51, no. 50, 2018.

[18] M. Suri *et al.*, Proc. IEEE International Electron Devices Meeting (IEDM), pp. 4.4.1-4.4.4, 2011.

[19] P. Hasler *et al.*, in Proc. NIPS, pp. 817-824, 1994.

[20] C. Diorio *et al.*, IEEE Transactions on Electron Devices, vol. 43, no. 11, 1996.

[21] B. W. Lee *et al.*, IEEE Transactions on Circuits and Systems, vol. 38, no. 6, 1991.

[22] Y. Horio *et al.* IEEE International Symposium on Circuits and Systems, 1990 (DOI: 10.1109/ISCAS.1990.112638)

[23] Y. Nishi ed. Advances in non-volatile memory, Elsevier 2014.

[24] H. Choi *et al.*, IEEE Transactions on Electron Devices, vol. 65, no. 1, 2018.

[25] G. Burr *et al.*, IEEE Transactions on Electron Devices, vol. 62, no. 11, 2015.

[26] P. Chen *et al.* IEEE International Electron Device Meeting, 2017 (DOI: 10.1109/IEDM.2017.8268337).

[27] T. Gokmen and Y. Vlasov, Frontiers in Neuroscience vol. 10, no. 333, 2016.

[28] A. Hirohata *et al.*, IEEE Transactions on Magnetics, vol. 51, no. 10, 2015.

[29] D. J. Newman *et al.*, "UCI repository of machine learning databases", Department of Information and Computer Science, University of California Irvine, CA, 1998.

[30] U. Saxena, D. Kaushik, M. Bansal, U. Sahu and D. Bhowmik, IEEE Transactions on Magnetics, vol. 54, no. 11, 2018..

[31] A. Sengupta *et al.*, Proc. IEEE Biomedical Circuits and Systems Conference (BioCAS), no. 7833852, 2016

[32] C. S. Thakur *et al.*, IEEE Trans. on Circuits and Systems- I, vol. 63, no. 2, 2016.

[33] A. Tripathi *et al.*, Proc.in IEEE International Symposium on Circuits and Systems (ISCAS), 2019.

[34] P. U. Diehl and M. Cook, Front. Comput. Neurosci. vol. 9, no. 99, 2015.

[35] G. Q. Bi and M. M. Poo, J. Neurosci. 18, pp. 1046410472, 1998.

[36] M. Pfeiffer *et al.* Frontiers in Neuroscience vol. 12, no. 774, 2018.

[37] P. Narayanan *et al.* IBM Journal of Research and Development, vol. 61, no. 4/5, 2017.

[38] A. Sengupta, A. Banerjee, and K. Roy, *Phys. Rev. Appl.* vol. 6, no. 064003, 2016.

[39] T. Bhattacharya *et al.* Proc. in IEEE International Joint Conference on Neural Networks (IJCNN), no. 18181462 (2018)

[40] C. Hu, "MOS Transistor in Modern Semiconductor Devices for Integrated Circuits", Pearson, 2010.

[41] S. Haykin, "Neural Networks," Pearson, 1994

[42] B. D. Yammenavar *et al.*, International Journal of Artificial Intelligence and Applications (IJAIA), vol. 2, no. 3, 2011.

[43] G. Khodabandehloo *et al.*,IEEE Trans. on VLSI Systems, vol. 20, no. 4, 2012.

[44] D. Neamen *et al.*, Electronic Circuits- Analysis and Design, McGraw Hill, 2006.

[45] AN-1515 "A Comprehensive Study of the Howland Current Pump" (PDF) (PDF). Texas Instruments, Inc. 2013.

[46] T. Gokmen *et al.*, Front. Neurosci. vol. 10, no. 333, 2016.

[47] S. Kim *et al.*,arXiv:1706.06620, 2017

[48] A. Biswas *et al.*, arXiv:1612.02233, 2016.

[49] D. Bhowmik *et al.*, *J Magn Magn Mater.,* vol. 489, no. 165434, 2019.

[50] S. Zhang *et al.*, Adv. Electron Mater., no. 1800782, 2019.

[51] P. Chen *et al.*, IEEE Trans. Electron Devices, vol. 62, no. 12, 2015.

[52] C. Li *et al.*, Nature Comm. vol. 9, no. 2385, 2018.

[53] S. Emori *et al.*, Spin Hall torque magnetometry of Dzyaloshinskii do- main walls, Physical Review B, vol. 90, no. 184427, 2014.

[54] S. Emori, U. Bauer, S.M. Ahn, E. Martinez and G. S. D. Beach, Nature Materials, vol.12, pp. 611616, 2013.

2019 20th International Conference on Research and Education in Mechatronics (REM)
May 23-24, 2019, Wels, Austria

Optimal Q-Learning Approach for Tuning the Cavity Filters

Even Sekhri
Department of Electrical Power Engineering and Mechatronics
Tallinn University of Technology
Tallinn, Estonia
even.sekhri@taltech.ee

Mart Tamre
Department of Electrical Power Engineering and Mechatronics
Tallinn University of Technology
Tallinn, Estonia
mart.tamre@taltech.ee

Rajiv Kapoor
Department of Electronics and Communication
Delhi Technological University
New Delhi, India
rajivkapoor.dtu@gmail.com

***Abstract*—The current research paper elucidates the optimal approach for tuning the Microwave Cavity Filters. The proposed solution uses Q-learning approach which is a special case of Temporal Difference used in Reinforcement Learning. The results are optimized using Lagrangian Multiplier. The proposed algorithm is tested on a commercially used filter. In this research work, only four screws were used for training the algorithm and for testing it. The algorithm could understand the strategies and could tune the reflection characteristics of the considered filter in 43 steps which proves the effectiveness of the algorithm to assist in the tuning process.**

Keywords— Q-Learning, Microwave Cavity Filter Tuning, Lagrangian Multiplier, Optimization

I. INTRODUCTION

Due to the recent rise in cellular communication, the network providers are motivated to install new Radio Base Stations (RBS) to manage the data traffic and to provide better connectivity. The spectrum is crowded now, and the need to filter out the desired frequency is high. Microwave (MW)/Radio frequency (RF) cavity filter is one of the major components in the RBS because of its High-Q factor, stability against environmental effects, reliable operation and robustness. To meet the connectivity requirements, raised by the substancial increase in the number of users, numbers of installed filter units have increased exponentially. As a consequence, the need for having high volume production of these filters is all-time high.

After the production, filters are assembled and pre-tuned. Due to various imperfections in the design, filter tolerances and material's dielectric constants, the filter needs to be properly tuned. Currently, tuning is done manually by the trained technicians and is a highly iterative, time-consuming and monotonous process. This process may vary from 15 minutes to a few hours by studying the scattering parameters (also called as S-parameters) response of the filter [1]. Attaining ample proficiency is also equally time-consuming even if some guiding documents are available in [2]. While rotating the screws, the technicians observe the S-parameters response of the filter on the Vector Network Analyzer (VNA) with the aim of tuning the filter in the well-defined frequency limits. Lot many filter topologies, with different frequencies and bandwidth, are available in the market which makes generalization next to impossible.

At times, even after being trained for a few months, some trainees remain inefficient in performing the task. The sustainability of the tuners also poses a big concern especially because finding new tuners is not a cakewalk. The major reason why tuning is difficult is because there is a complex correlation between the tuning elements and the filter response [3]. The beforementioned paragraphs suggest that tuning is a cumbersome task. Hence, searching automated means of doing this process is the bone of contention of the researchers in the field to remove the bottleneck of production.

The next section discusses all the potent efforts made by the researchers for the advancement in the field.

II. LITERATURE SURVEY

Automating the cavity filter tuning by using Machine Learning (ML) has been studied quite exhaustively by the researchers for decades. The concept of using ML Algorithm for finding the solution to this tuning problem was first presented in 1989 [4], in which the researchers did not use any Mathematical Modeling or Coupling Matrix Extraction procedure. They used waveguide filter and presented an algorithm that could only assist the human operator to fast tune but could not tune the filter automatically. The major drawback of their time was the assumption that the coupling screws are already correctly pre-adjusted. Continued researches in the area many other techniques were proposed. These included algorithms based on Adaptive Models [5] which was based on the approximate network models that were derived using the prototype network. This is not feasible to implement in real life. Fuzzy Logic was also implemented in this area [6]–[10] which although gave more flexibility, but no consideration was given to human experience in [6] and [7]. The usage of inevitable human experience came only in [8][9] and [10] which could assist in the accurate tuning. The major issue with the methods used by these researchers was that they were based on coupling matrix extraction from the mathematical model. However, the accurate mathematical model for the cavity filter is not reported into the literature yet. Another major issue with these researches was related to having an insufficient amount of linguistic rules. Both these factors limited this method to

978-1-5386-9257-8/19/$31.00 ©2019 European Union

be used in real life. Artificial Neural Networks (ANN) was used to get the inverse model of the filter [11]–[14]. This method was also not without its shortcomings as it required huge amount of training data, which further led to increased time in the filter tuning. Later, a hybrid approach of combining two or more techniques was proposed in [15]. In recent times, researchers are resorting to Reinforcement Learning (RL) [16] in which they had used only two screws for training their model. Selection of the reward cost was their key concern. The same authors also tried a vector field approach [17] but the algorithm was not at all intelligent. Later in [18] the authors combined the techniques in [16] and [17] and used DDPG framework to have continuous RL but used a simple filter model. They turned only four screws that too in a limited range, with the curve that chaged only on the vertical side.

The above overview indicates the amount of efforts put forth by the researchers for the advancement in the area. However, the findings of these numerous analytical and practical approaches were either incomplete or deficient or nonviable; as none of the researches were able to find the solution that could tune the complex filters automatically. The reason why most of the methods proposed by researchers cannot be implemented in real life is that they modelled the filters only but did not focus on the role of human intelligence. The role of an experienced tuner is inevitable. The data from the human experience can exponentially increase the performance of the system. This approach has been discussed in recent times in [10] and [16] but, yet, it needs to be explored further to achieve Automated Filter Tuning.

Developing a meaningful algorithm can reduce the time and fabrication costs significantly. Computerized Automatic Tuning (CAT) of MW filters is challenging but it indeed will meet the demands of rapidly increasing market [19][20][21]. It will have the potential of not only tuning the filter automatically and reduce human dependency but, it also can open the path for understanding the patterns behind this process. The ideal automated CAT must not have any human involvement; mathematical filter modeling should be avoided; and finally, the filter must reach its optimum state of tuning. Considering all these requirements, and understanding the fact that the humans train their Natural Neural Network while they learn and try to tune the filter makes the usage of ANN the basis of this research work. Reinforcement Learning (RL) does not need any supervision rather it learn the patterns from the environment in which they are implemented. In this research work, RL has been used and the Reflection Curve (S_{11}) is considered. Q-Learning is used in this research work and the Loss function is optimized using Lagrangian Multiplier to obtain the Global Minima. The subsequent section illustrates the Q-Learning approch and discusses the Learning Model used in this current research work which is then optimized.

III. Q-Learning Algorithm Based Tuning

A. Preliminaries

The Device Under Test (DUT) is a cavity filter which is connected to Vector Network Analyzer (VNA) using input/output ports. VNA is the device which provides input to the filter and then also depicts the current frequency response of the filter in the form of S-parameters curves. The data can be extracted in terms of Insertion Loss (the useful energy dissipated as heat), Return Loss (the signal reflected back towards the source), Passband Signal (the absolute value should be as close to 0 dB as possible) and the Band-Reject Signal (with or without Transmission Zeros). In this research work, we considered only Return Loss Curve (S_{11}) as the state. The goal of the tuning is only to minimize the absolute value of the Return Loss and to ensure it is below the target value of -16dB in the passband zone.

B. Formulation

Markov Decision Process (MDP) provides a framework for mathematical modeling when outcomes are partly under the decision maker's control and partly random. In [16], researchers used Q-Learning in the field of Filter Tuning which was based on Stochastic Gradient Descent (SGD) Algorithm. The filter model used by them was very simple and often faced the problem of trapping into local minima. In this work, authors decided to use Q-learning approach considering Convolutional Neural Network (CNN) architecture so as to reach towards the Global Minima of the defined Cost Function using Lagrangian Multiplier. In MDP, the Learner and the Decision-Maker is named as "Agent" and it interacts with the "Environment". MDP is basically defined by *s,a,r,p* and γ where:

$s \in S$ – the State of the considered system

$a \in A$ – the Action taken by the agent

$r : S \times A \rightarrow R$ – the Reward Function

$p(s_{future} \mid s_{now}, a_{now})$ – the state transition (probability) function or system dynamics

γ – the discount factor $\in (0,1]$

In MDP (and RL), the policy π is defined as a function which specifies the action to be taken corresponding to each state i.e. $\pi : S \rightarrow A$. At any instance the agent is in the state $s \in S$, action $a \in A$ and then its future state is s_{future}. This has been explained pictorically in Fig. 1 below.

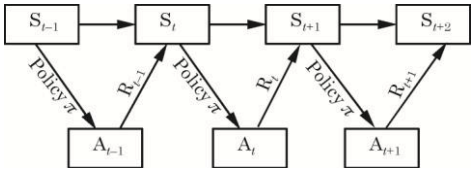


Fig. 1: General Sequence in RL

In our case, agent is the designed algorithm; states are the points considered on S-parameter curve; actions are clockwise and anti-clockwise turning angles of the screws; the discount factor γ is set to 0.95 in this research. The target of the Q-Algorithm is to maximize the future rewards after learning the policy $\pi(s_{now}) \rightarrow a_{\text{now}}$. In this research work, the target of the Q-algorithm is to ensure that the return loss i.e. S_{11} parameters of the filter is below the target value which is -16 dB in our case. The s-parameters of the DUT are obtained and are compared with the target value in the passband zone. In case, the S_{11} parameters of DUT are already below the target line, then, the distance is defined to be zero. But, in alternate case, the distance d is recorded by taking the absolute value of the difference between the two curves. If this d value is getting smaller, the system receives a reward point of 1 else it gets a zero as its reward. When the system is presented with a state, the system takes an action against it and receives the corresponding reward and reaches the next (future) state and saves this sequence. After this, a mini-batch of the saved sequence is extracted to train the designed network. Further, the future action a_{future} which actually maximizes the Q-value of state s_{future} is selected for the future cycle.

C. The Q-Learning Model

The quality i.e. Q-function (using Bellman's Equation) is given by (1) below.

$$Q(s_{now}, a_{now}) = \underbrace{r_{\text{now}}}_{\text{Immediate reward}} + \underbrace{\gamma \max_{a_{\text{future}}} Q(s_{\text{future}}, a_{\text{future}})}_{\text{Discounted/Future Reward}} \quad (1)$$

Keeping optimal value V* (stated in (3)) then an optimal policy is to decide deterministically, and the optimal action is given by (2).

$$a^*(s_{\text{now}}) = \arg\max_{a_{\text{now}}} Q^*(s_{\text{now}}, a_{\text{now}}) \quad (2)$$

(where a^* is the optimal action and Q^* is the optimal Q-value)

$$\text{i.e.} \quad V^*(s_{\text{now}}) = \max_{a_{\text{now}}} Q^*(s_{\text{now}}, a_{\text{now}}) \quad (3)$$

The Q-Algorithm used is drawn in Fig. 2. The update rule is defined by (4) and (5).

$$Q_{\text{est}}(s_{\text{now}}, a_{\text{now}}) \leftarrow Q_{\text{est}}(s_{\text{now}}, a_{\text{now}}) + \alpha\left[r_{\text{now}} + \gamma \max_{a_{\text{future}}} Q_{est}(s_{\text{future}}, a_{\text{future}}) - Q_{\text{est}}(s_{\text{now}}, a_{\text{now}})\right] \quad (4)$$

$$\& \quad s_{\text{now}} \leftarrow s_{\text{future}} \text{ until termination} \quad (5)$$

where: α = learning rate or step size (0,1].

To deal with non-exploration i.e. to update α, we used random exploration. At the stated probability of $\frac{1}{N}$ where N = number of trials i.e. number of times we have used "(s, a)-pairs", the program took random action instead of optimal one and later, the probability of α was lowered automatically.

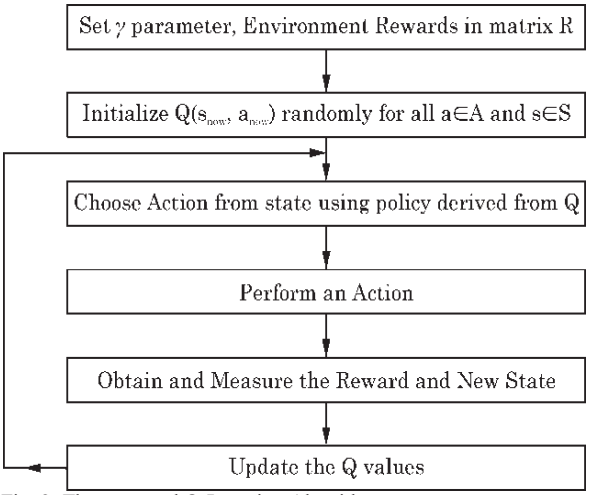


Fig. 2: The proposed Q-Learning Algorithm

The loss function is defined by (6) as:

$$\mathcal{L}(w) = \mathbb{E}\left[\left(\underbrace{r_{\text{now}} + \gamma \max_{a_{\text{future}}} Q_{\text{est}}(s_{\text{future}}, a_{\text{future}}, w)}_{\text{target}} - \underbrace{Q_{\text{est}}(s_{\text{now}}, a_{\text{now}}, w)}_{\text{estimated}}\right)^2\right] \quad (6)$$

Then, the Q-Learning gradient is computed by (7) as:

$$\frac{\partial \mathcal{L}(w)}{\partial w} = \mathbb{E}\left\{\left(r_{\text{now}} + \gamma \max_{a_{\text{future}}} Q_{\text{est}}(s_{\text{future}}, a_{\text{future}}, w) - Q_{\text{est}}(s_{\text{now}}, a_{\text{now}}, w)\right)\frac{\partial Q_{est}(s_{now}, a_{now}, w}{\partial w}\right\} \quad (7)$$

To remove correlation, we built the data set from agent's own experience and we took the following action:

1. Taking action a_{now} using greedy policy
2. Store the transition T $(s_{now}, a_{now}, r_{future}, s_{future})$ in replay memory $\mathcal{D}$.
3. Taking sample random mini-batch from the stored transition
4. Optimizing MSE between Q-network and Q-learning targets given by (8)

$$\mathcal{L}(w) = \mathbb{E}_{s_{\text{now}}, a_{\text{now}}, r_{\text{now}}, s_{\text{future}} \sim \mathcal{D}}\left\{\left(\underbrace{r_{\text{now}} + \gamma \max_{a_{\text{future}}} Q_{\text{est}}(s_{\text{future}}, a_{\text{future}}, w)}_{\text{target}} - \underbrace{Q_{\text{est}}(s_{\text{now}}, a_{\text{now}}, w)}_{\text{estimated}}\right)^2\right\} \quad (8)$$

The loss function is optimized end-to-end by Lagrangian optimization mentioned in (9) i.e.

$$\mathcal{L}(f, \gamma, x) = f(t, y) + \gamma . x(t, y) \quad (9)$$

and, later we equated $\frac{d\mathcal{L}}{dx} = 0$ to find the global optimum value with the constraint given in (10) i.e. :

Loss Function > Threshold Value (10)

The next section of this research paper is dedicated to the experimentation methodology used and the results obtained henceforth. A detailed discussion of the results is also presented.

IV. Experimentation

A. General Implementation Details

In this research work, the DUT is a version of Huawei Cavity Filter with two channels. We used only one channel i.e Receiver Side (RX Side) and four resonator screw of RX Side for performing the actions. Hence, we had eight actions to be performed (rotation angles of all the four screws in both directions by one full rotation i.e. 360°). The environment used in this research work is simulated, where we generated training samples by randomly inserting the screws one by one manually.

As a function approximator, we used CNN Architecture where the input is the passband section of the S-parameters. The input of Q-network was set at 25 artificial neurons and the output layer consisted of eight neurons indicating the eight valid actions for the Q-values.

For training the network, we randomly drew 35 sequences from the collected database $\mathcal{D}$. A greedy policy was used so as to obtain the biggest estimated Q-value.

B. Results

We trained the defined network for 150 epochs each with 1000 tuning steps as the maximum value. After training the network, it was tested with 50 random states. After 100 training epochs, the algorithm was tested with 100 random states.

Fig. 3 provides the results of our algorithm. It was found that after training the network it could successfully drive the curve below –16 dB return loss target and just after 43 steps the filter was successfully tuned. Global Minima was ensured by the use of Lagrangian Multiplier.

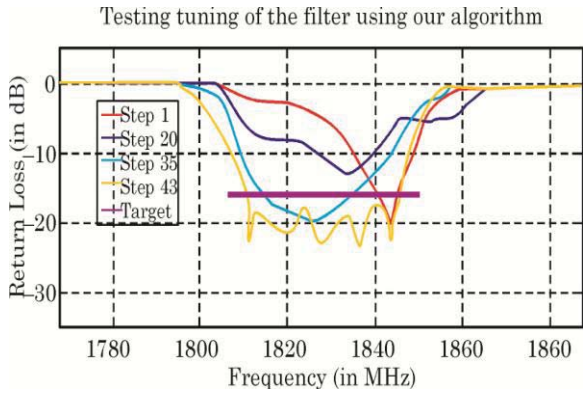


Fig. 3: The experimentation results

C. Discussion of the Results

The current research proved the effectiveness of Q-learning, the proposed algorithm could successfully find the correct screw position and global optimum. The major limitations and hence, further implications of our research include:

a) In this present research, only reflection characteristics are considered without considering the forward insertion loss.

b) Further, only the distance between the current reflection characteristics and the target value within the passband was used. The limit lines on the sides (for sharper transition between stopband and passband), however, were unobserved.

c) Our reward policy considers only the distance between current and target values, the penalty for complete removal or insertion of the tuning screw, however, is not considered. This is a major consideration in real scenarios and hence must be given a heavy penalty. In addition, no reward was assigned to successfully tuning the filter.

V. Conclusion

The current research highlighted the difficulties of manual tuning of cavity filters which led to the need of tuning the filters by autonomated means. A Q-learning algorithm was propsed as a viable solution. The designed algorithm could learn the tuning strategies and updated the estimated Q-values iteratively. The effectivess of the algorithm was ensured by experiments. The designed algorithm could successfully tune the raw filter (using only four resonator screws) after 43 steps.

References

[1] T. Kacmajor, P. Kant, and J. J. Michalski, "Microwave filter tuning for different center frequencies based on Artificial Neural Network and phase compensation," *2014 20th Int. Conf. Microwaves, Radar Wirel. Commun. MIKON 2014*, 2014.

[2] A. Lindner, E. Biebl, and A. Strasse, "A Manual Tuning Method for Coupled Cavity Filters," in *Proceedings of 36th European Microwave Conference*, pp. 1340–1342.

[3] L. Han, K. Wu, X.-P. Chen, and F. He, "Influence and Tuning of Tunable Screws for Microwave Filters Using Least Squares Support Vector Regression," *Int. J. RF Microw. Comput. Eng.*, vol. 20, no. 4, pp. 422–429, 2010.

[4] A. R. Mirzai, C. F. N. Cowan, and T. M. Crawford, "Intelligent alignment of waveguide filters using a machine learning approach," *Microw. Theory Tech. IEEE Trans.*, 1989.

[5] P. Harscher, J. Hofmann, R. Vahldieck, and B. Ludwig, "Automatic Computer-Controlled Tuning System for Microwave Filters," in *30th European Microwave Conference*, 2000, pp. 1–4.

[6] V. Miraftab, S. Member, and R. R. Mansour, "Computer-Aided Tuning of Microwave Filters using Fuzzy Logic," in *Miraftab, Vahid Mansour, Raafat R*, 2002, vol. 50, no. 12, pp. 2781–2788.

[7] V. Miraftab and R. R. Mansour, “A robust fuzzy-logic technique for computer-aided diagnosis of microwave filters,” *IEEE Trans. Microw. Theory Tech.*, vol. 52, no. 1 II, pp. 450–456, 2004.

[8] V. Miraftab and R. R. Mansour, “Tuning of microwave filters by extracting human experience using Fuzzy Logic,” *IEEE MTT-S Int. Microw. Symp. Dig.*, vol. 2005, no. C, pp. 1605–1608, 2005.

[9] V. Miraftab and R. R. Mansour, “Automated microwave filter tuning by extracting human experience in terms of linguistic rules using fuzzy controllers,” *IEEE MTT-S Int. Microw. Symp. Dig.*, pp. 1439–1442, 2006.

[10] V. Miraftab and R. R. Mansour, “Fully automated RF/microwave filter tuning by extracting human experience using fuzzy controllers,” *IEEE Trans. Circuits Syst. I Regul. Pap.*, vol. 55, no. 5, pp. 1357–1367, 2008.

[11] J. J. Michalski, “Artificial Neural Networks Approach in Microwave Filter Tuning,” *Prog. Electromagn. Res. M*, vol. 13, pp. 173–188, 2010.

[12] J. J. Michalski, T. Kacmajor, J. Gulgowski, and M. Mazur, “Consideration on Artificial Neural Network Architecture in Application for Microwave Filter Tuning,” *Piers Online*, vol. 7, no. 3, pp. 271–275, 2011.

[13] J. Gulgowski and J. J. Michalski, “The analytic Extraction of The Complex-Valued Coupling Matrix and its Applcation in the Microwave Filter Modeling,” in *Progress In Electromagnetics Research*, 2012, vol. 130, no. July, pp. 131–151.

[14] J. J. Michalski, “Inverse Modeling in Application for Sequential 0Filter Tuning,” in *Progress In Electromagnetics Research*, 2011, vol. 115, no. February, pp. 113–129.

[15] T. Kacmajor and J. J. Michalski, “Neuro-fuzzy approach in microwave filter tuning,” *IEEE MTT-S Int. Microw. Symp. Dig.*, pp. 2–5, 2011.

[16] Z. Wang, J. Yang, J. Hu, W. Feng, and Y. Ou, “Reinforcement learning approach to learning human experience in tuning cavity filters,” *2015 IEEE Int. Conf. Robot. Biomimetics, IEEE-ROBIO 2015*, pp. 2145–2150, 2015.

[17] Z. Wang, S. Jin, J. Yang, X. Wu, and Y. Ou, “Real-time tuning of cavity filters by learning from human experience: A vector field approach,” *Proc. World Congr. Intell. Control Autom.*, vol. 2016–Septe, pp. 1931–1936, 2016.

[18] Z. Wang, Y. Ou, X. Wu, and W. Feng, “Continuous Reinforcement Learning With Knowledge-Inspired Reward Shaping for Autonomous Cavity Filter Tuning Continuous Reinforcement Learning With Knowledge-Inspired Reward Shaping for Autonomous Cavity Filter Tuning*,” in *Proceedings of 2018 IEEE International Conference on Cyborg and Bionic Systems*, 2018, no. December, pp. 53–58.

[19] H. Hsu, S. Member, H. Yao, S. Member, and K. A. Zaki, “Computer-Aided Diagnosis and Tuning of Cascaded Coupled Resonator Filters,” in *IEEE Transactions on Microwave Theory and Techniques*, 2002, vol. 50, no. 4, pp. 1137–1145.

[20] J. B. Ness, “A unified approach to the design, measurement, and tuning of coupled-resonator filters,” *IEEE Trans. Microw. Theory Tech.*, vol. 46, no. 4, pp. 343–351, 1998.

[21] G. Pepe, F. Gortz, and H. Chaloupka, “Computer-aided tuning and diagnosis of microwave filters using sequential parameter extraction,” *Microw. Symp. Dig. 2004 IEEE MTT-S Int.*, vol. 3, p. 1373–1376 Vol.3, 2004.

International Journal of Research in Engineering, Science and Management
Volume-2, Issue-6, June-2019
www.ijresm.com | ISSN (Online): 2581-5792

Problems of Scaling and Corrosion in Geothermal Water: Indian Scenario

Shivam Sharma[1], Sarthak Aggarwal[2], Yash Kumar Baghel[3], Ajit Kumar[4], J. P. Kesari[5]
[1,2,3,4]*B.Tech. Student, Dept. of Mechanical & Automobile Engg., Delhi Technological University, Delhi, India*
[5]*Associate Professor, Dept. of Mechanical & Automobile Engg., Delhi Technological University, Delhi, India*

***Abstract*: India is looking at tapping the geothermal resources that are available at majorly seven regions in India. India is ranked low in terms of Geothermal potential with low/medium heat enthalpy but there are enough resources with temperatures in the moderate to high range to generate power through hybrid systems if not independently through geothermal. There are a few geothermal power plants in the pipeline but nothing significant on the ground as yet. Since the water from geothermal sites is not chemically inert or pure, there is always the problem of scaling and corrosion in the power plant setup. Silica scaling has long been a hurdle in efficient power extraction from geothermal resources. In relation to this problem, data from all of these seven Indian geothermal provinces was extensively analyzed to draw conclusions on the locations which are the most feasible locations for setting up power plants based on the comparison of several chemical factors such as hardness, pH and composition of water from various sites. Although this is just a suggestive feasibility based on chemical aspect of thermal fluids, yet this aspect is a vital one. Also, the scope of chemical extraction from these sites is discussed. Extracting the minerals from the geothermal fluid would not just safeguard the plant from the problem of scaling and corrosion but also reduce the cost of power generation by contributing to the profitability.**

***Keywords*: Geothermal Energy, Green energy, Corrosion, Extraction, Scaling.**

1. Introduction

Today, when countries round the globe are looking to meet the ever increasing demands of their energy for their residents, the need to shift towards newer and green renewable resources can be denied by none. If these new resources are also capable of somewhat making the population self-sufficient in meeting their energy needs, there cannot be a better blessing than this for the countries like India where the gap between energy demand and supply is widening by the hour. This is probably the biggest reason and benefit behind the success of solar energy in India because it makes a common man a little independent by generating own electricity at home. One such resource can be Geothermal which is still widely untapped on a global scale. Geothermal Energy is the renewable underground energy which is available in the form of heat. This won't provide every commoner the power to drill at his home and use geothermal water for power generation because this comes with its set of foundations as well. Common man cannot generate electricity through this but communities can manage to set up infrastructures to use the geothermal fluids to meet thermal domestic needs but only if it is geographically viable at their location. Most vital advantage of this resource over other resources is its availability all-round the year independent of the weather. It has been estimated that heat energy within the 10,000 metres of earth's surface is 50,000 times more than all the oil and gas resources worldwide [1].

The temperature, composition of fluid, pressure, mass flow rate and many other factors vary from place to place like for every other natural resource. Depending upon these factors, the utility of geothermal resource varies from direct domestic heating to power generation to mineral extraction. Low temperature fluids can be used for domestic purposes, agriculture, horticulture etc. Medium temperature range fluids of around 300K can be used for power generation while the higher temperature range fluids can be used for industrial applications as well. Geothermal fluids can also be seen as a source for extraction of some metals and minerals. Since this fluid comes in contact with various rocks and ores inside the earth's surface and that too under high pressure and temperature conditions, the fluid is bound to pick up minerals and metals in some or the other form from there. This leads to the unavoidable menace of scaling and corrosion of all the pipelines and tanks that the geothermal fluid comes in contact with. One of the principle conditions of extraction is the economical side of it, the process and the end result that is the extracted mineral should be economical to extract from this water. This would require a balance between low concentration of the mineral and high cost of extracting it. Only when both of these factors are worked out to provide a profitable layout can the extraction be done. With time, the cost of power generation through geothermal sources is also coming down and becoming more and more competitive. This also makes this resource a promising as well as a fulfilling one.

2. Review of literature

Cinti et al. in their paper discuss about the geochemistry of several areas in India and they concluded that the surface temperatures in Himalayan region are generally on the higher side, it being higher in the Parvati Valley than the Beas Valley. While circulating at depths inside earth's surface, the water

picks up saline nature due to its interaction with rocks. The relatively high salinity of Sohna spring points to longer residence time or circulation in crust. Dissolved hydrogen in huge amounts was also found in roughly all the samples assessed from Parvati, Beas and Sohna region which pointed towards fault activity [2].

Corrosion rates were assessed by Cabrini et al for various chemical compositions over a period of 7 months of exposure. Initial exposure led to the formation of layer of corrosion products for a brief period at a high rate which tends to decrease for subsequent exposure. The layer acted as protective inert layer for a short period before flow condition overcome it to increase the corrosion rate again. As the scales thicken their adhesive properties decreases and hence the rate of corrosion increases rapidly in the longer run [3].

Discussing about the scaling problem in Geothermal heat systems, K. Rafferty talks about water chemistry and how scaling is often overlooked in the GHP operation. For the open loop usage of heat pump, scaling can be a common problem as the water is directly used repeatedly. But in the case of closed loop, concern shifts towards the desuperheater. Hardness and alkalinity can give clear indications about the chance of scaling. Most common form of scale is that of Calcium Carbonate which can be attributed to the groundwater that is used. Carbonate Hardness that includes the carbonates and bicarbonates of Calcium and Magnesium then there is non-carbonate hardness which can be characterized by comparatively greater solubility. It is the Carbonated Hardness that is a major scaling promoter. Alkalinity is a measure of how able the water is in neutralizing the acid. Water quality problems also increase with the subsequent increase in Total Dissolved Solids (TDS). Two commonly used indices in the water treatment applications are Ryznar Stability Index (RSI) and Langelier Saturation Index (LSI). Since both these indices are based on the saturation point of carbonates, they provide a better image of scaling than corrosion. Lime deposition increases with the increase in water usage and it also increases with the increase in temperature of the water. Scale formation as indicated by LSI, increases with increase in pH and is evident beyond a pH of 7.5. A scale formation of .03 inches on heat exchanger system can increase the power consumed by 19% at the operation conditions of 55°F in GHP. To qualitatively determine the degree of scaling, we need the information of several factors such as TDS, pH, calcium hardness, and temperature. pH determination should be treated as primary indicator and if it is found to be greater than 8, relevant data should be gathered to determine RSI and LSI. Weak acid can be circulated through heat pump to remove the scale periodically [4].

Pierre Ungemach discusses the solutions to this problem of scaling and how it can be handled better for low enthalpy Geothermal Environments. Damage diagnosis is the first step in recognizing the problem and then addressing it. It can be diagnosed in several ways. One such diagnosis is through the assessment of pressure and flow characteristics. Others are direct damage assessment through inspection by ultrasonic, production tools and multifinger calipers. The scale removal is conventionally achieved by cleaning the well through rockbits which are drill driven or hydrojetting tools. Weak acids can also prove to be an efficient damage remedy for this scaling problem when dilute acid is periodically injected into the system. The chemical inhibitors can also be used to mainly address two concerns that is to prevent corrosion and attack scales. Sulphate reducing bacteria promotes microbiological corrosion which is a serious cause of damage to injector wells in particular.[5]

3. Chemistry of geothermal fluids

Geothermal fluids constitutes of a wide variety of dissolved chemical species in different concentrations. Geothermal fluids can be characterized by frequently cited parameters such as TDS (Total Dissolved Solids) and pH of the fluid. TDS or Total Dissolved Solid is measured in parts per million (ppm) or milligram per litre (mg/L) and it gives the estimation of the amount of chemical salts dissolved in the geothermal fluid. The TDS value of geothermal fluid varies from few hundred to more than 3,00,000mg/L, whereas in high-temperature geothermal regions the TDS lies between 6,000 to 10,000 mg/L. In geothermal fluids, the nature and quantity of dissolved chemical salts and species depend upon the local earth's structure or geology and the temperature of geothermal fluid. Generally, it is noted that the amount of dissolved solids in low-temperature resources is lower than that of dissolve solids in high-temperature geothermal resources. Another parameter pH denotes acidity or basicity of the geothermal fluid. The acidic or corrosive fluids have a pH of less than 7, and the alkaline fluids have a pH of greater than 7. In the case of neutral fluids, their pH should exactly be equal to 7 at room temperature. Mainly the pH of geothermal resources lies between moderately acidic (5.5) and moderately alkaline (8.5) [6]. In low temperature geothermal water the dissolved solids majorly composed of silica (SiO_2), calcium (Ca), sodium (Na), potassium (K), chlorine (Cl), sulphate (SO_4), and bicarbonate (HCO_3). The geothermal fluid also contains various elements such as mercury (Hg), fluorine (F), boron (B), and arsenic (As) in small amounts [7]. Arsenic, being toxic if present in higher concentrations is of environmental concern and it has to be eliminated before extraction of other materials. In geothermal resources the gases are also dissolved in the fluids. Majorly dissolved gases include ammonia (NH_3), carbon dioxide (CO_2), hydrogen sulphide (H_2S), and methane (CH_4) [7]. Hydrogen sulphide dissolved in geothermal fluids is toxic and dangerous for animals as well as humans. As the geothermal fluid flow it interacts with surrounding rocks and reacts chemically to form chemical compounds or minerals. Carbonates, sulphates, and silica are produced as a result of these complex chemical reactions. Some of the minerals either dissolved fully or some of them are precipitated from solution and in some cases even substitution of elements within a mineral takes place. These changes may or may not attribute to volume changes of the

reservoir rocks, which means the porosity and the permeability of the rock may or may not be affected due to these chemical reactions [6]. If we consider the case of silica and calcite, the solubility of silica in geothermal fluid decreases with decrease in temperature of fluid whereas the pressure change has negligible effect on solubility of silica. So, SiO_2 generally precipitates due to abrupt change in temperature of subsurface, in pores or fracture in the rocks. But the case of other minerals like calcite (calcium carbonate, $CaCO_3$), dolomite (magnesium carbonate, $MgCO_3$), and sulphate species is totally opposite i.e. these minerals have higher solubility at lower temperatures than at high temperature [6].

In above paragraph, we came to know about different chemical compounds and gases which are dissolved in geothermal fluid. Upon exploitation the fluid undergoes thermodynamic changes which results in scaling, corrosion of surface in contact with fluid and sometimes environmental problems of liquid disposal and emission of toxic gases like H_2S [8]. Scaling can be described as a process of accrual or collection of undesirable material on the surfaces which interact with the geothermal fluid. The scaling is common phenomenon and can be observed in geothermal power plants and also in reservoir surroundings. It is a very rapid, rampant, and complex process. It depends upon various factors such as pressure and temperature of geothermal fluid, operating conditions and time of interaction between fluid and rock. When the geothermal fluid boils, dissolved gases are released due to which pH level of fluid increases and it becomes more alkaline due to which metal sulphides and calcites are formed. These minerals form hard scales on inner portions of surface pipes and well casing [9]. Normally crystallization scaling takes place in geothermal installations when the ionic product of less soluble salt becomes greater than its equilibrium solubility product [8]. Common forms of scale include carbonate minerals ($CaCO_3$), metal oxides, amorphous silicates (SiO_2), and sulphides. The most profound geothermal scales are of calcite ($CaCO_3$), and silica (SiO_2). The colour of both these scales is white and these cannot be distinguished by naked eyes. Sometimes due to corrosion in geothermal fluid pipe, iron sulphide is formed due to which the silica scales tends to appear grey or black [10]. To estimate the scaling potential of the geothermal fluid its TDS, pH, temperature, total hardness, and total alkalinity should be known. Usually the scaling problems are common for fluids whose pH value is above 7.5 [6]. Various methods are available to prevent and control scaling in geothermal systems. The most used methods include adding of scaling inhibitors, acidifying geothermal fluid to maintain mineral equilibrium, proper construction of geothermal plant, and by removing scales chemically or mechanically [8].

Another harmful effect of geothermal fluid is corrosion of systems used for handling fluid and also of the equipment used for utilization and exploitation of geothermal fluid [8]. The effects of corrosion of geothermal fluid are influenced by the chemical composition of fluid. The qualitative features of geothermal fluid such as lower pH, high concentration of salts and dissolved gases increase the corrosion potential. The geothermal fluid contains a wide variety of dissolved components varying from halogen acids to strongly sulphur containing acidic fluids to neutral or alkaline pH water, which effectively corrode most of alloys and even remove the protective layer formed due to scaling of silica, calcite or metal oxide. The chemicals species which are present majorly and make geothermal fluid highly corrosive are chlorides (Cl^-), sulphides (HS^-), sulphate (SO_4^{2-}) and bicarbonates (HCO_3^-). Further presence of oxygen, hydrogen ions, and fluorides supplements the corrosion of metals [8]. The rate of corrosion is enhanced if geothermal fluid is contaminated with oxygen. Other factor which affects rate of corrosion is release of dissolved gases because it is seen that as the dissolved carbon dioxide are released, the pH of geothermal fluid tends to increase. The rise in pH of fluid reduces the effectiveness corrosion action on steel. The flow conditions of geothermal water like turbulent flow and stagnant water conditions also increase the corrosion potential [6]. The different types of corrosions that are evident in geothermal installations are uniform corrosion, pitting, crevice, sulphide stress cracking, stress corrosion cracking, under deposit corrosion, galvanic corrosion and impingement. Components which are more sensitive towards corrosion are steel casings, screens and heat exchangers. In order to prevent corrosion the first step can be use of noble metals if there are no economic or technical constraints. Corrosion can also be limited by using corrosion resistant materials such as plastics or by not coagulating materials with distinguished electro-chemical potential. Coated pipes and casing can also prevent the problem of corrosion. Coating can be done by using thin plastic or epoxy layer on inner surface of pipes [6].

Some of the dissolved chemical components of geothermal fluid have a value and can be utilized if they can be easily extracted, separated, and purified. The valuable minerals can be recovered from geothermal resources, whether they are present in solid geothermal residues like scales and sludges or in the geothermal fluid. The process of extraction should be economical, profitable, and technically feasible, and the mineral should be available in abundance in the geothermal resource. The main chemical compounds extracted are silica, lithium, and alkali metals. The other subsidiary products that can be recovered are some inexpensive salts like NaCl, $CaCl_2$, and some other salts. Minerals like limestone and metals like gold and silver can also be extracted inexpensively [6]. The process used for the recovery of minerals from solid materials like sludges and scales is acid leaching and biochemical leaching. Acid leaching is generally used for extracting silica in pure form, and in this process hydrochloric acid is used to leach iron from the residues. Biochemical leaching is the process utilized for the recovery of metals from low-grade chalcopyrite and aurum ores. The process is carried out in the presence of acid-loving bacteria which leach out toxic metals. Sometimes

the residual products in this process are refined such that they can be used directly for further applications. Sorption, Evaporation, Precipitation as sulphides can be used to recover metals from geothermal water. Uranium, cobalt, zinc, manganese, and lithium ions are recovered by using synthetic resins, which adsorbs ions from the solution in the sorption process. Silica is very harmful to a geothermal resource as it causes scaling due to which geothermal resources cannot be utilized properly. In order to extract other minerals silica should be removed firstly. Silica can be removed in the form of quartz (SiO_2). Most of the geothermal resources have lithium in abundance, so these resources become a rich source of lithium extraction. By precipitation of lithium salts directly in the fluid lithium can be extracted. The main problem which arises during the extraction process is that despite the high concentration of precious metals in the resources it may not be convenient to extract these metals purely as the process involved will be complex, expensive, and risky.

4. Geothermal resources of India

Organized synergic research, development and corroboration program is being ventured by several organizations such as Geological Survey of India, National Aeronautic Limited, Bangalore, IIT Delhi, National Geophysical Research Institute (NGRI), ONGC, etc. As per the research analysis from different case studies and audits, locations of 350 inherent hot springs, scattered in around seven provinces have been determined all over the country. A persistent surface temperature range from 37°C - 90°C with an increasing surface discharge of 1000 l/m has been identified. Out of the water resources that have been identified a maximum number of them are of low temperature, which would be used in thermal applications. Only a certain number of locations are being identified where a spawning 10,000 MW of electrical power has been predicted. Small prototype plants have been made to utilize the geothermal resources to accumulate power from the resources present at Puga and Chumathang (Ladakh, J&K) and Manikaran (HP). The provinces are being divided as per their locations as The Himalayas, Cambay, Godavari, Sohana, Mahanadi, West coast, and Son-Narmada-Tapi (SONATA). The above provinces are affiliated with fissures or diminution tectonics where the flow of heat and the gradient of geothermal energy is optimum.

In some particular locations such as Puga valley of Ladakh and J&K and Tattapani (District Sarguja), Chattisgarh, these intermediate enthalpy geothermal reservoirs have been relevant in power generation with the help of a Binary Cycle structure. After an extensive research, a hydrothermal reservoir at a depth of 3 km at 260°C temperature in Tattapani has been determined by NGRI, Hyderabad.

To tap the inherent geothermal energy, the Ministry of New and Renewable Energy Resources have planned to develop the Puga valley in Ladakh (in March 2008 by Union Ministry of New and Renewable Energy). An estimated cost of the installation of the 300 kW electric power plant was around Rs 4.6 crore. Extensive research of Badrinath-Tapovan in Uttaranchal, Satluj-Spiti, Beas and Parbati valley in Himachal Pradesh and Surajkund in Jharkhand have been undertaken by NGRI, Hyderabad to get an optimized location to generate a Renewable source of Power and could cover only diminutive industries as most of these resources are situated in rural areas.

Fig. 1. Location of the seven geothermal provinces in India
(Source: indiaenergyportal.org)

In The Himalayan province, the main location is Puga valley (J&K), and Manikaran (HP) with a heat flow rate of 468 mW/m^2 and a thermal Gradient of 234 °C/km, a thermal gradient of 76-96 °C/km in Sohana province has been determined. The Main location in West coast province is Unai (Maharashtra) with a heat flow of 129 mW/m^2 and a thermal gradient of 59 °C/km. Tuwa (Gujarat) is identified as the main location in Cambay province with a heat flow rate of 93 mW/m2 and a thermal gradient of 70°C/km. The Son-Narmada-Tapi province is present on the Deccan Plateau with main locations as Jalgaon (Maharashtra), Tattapani (Chattisgarh), Bakreshwar (Bihar) with a heat flow rate of 120-260 mW/m^2 and a thermal gradient of 60-90°C/km. The heat flow rate of 104 mW/m^2 and thermal gradient of 60°C/km has been determined in the Godavari province in the last province Mahanadi a heat flow rate of 200 mW/m^2 and a thermal gradient of 90°C/km has been determined. All this extensive data research has been done by the National Geophysical Research Institute (NGRI), Hyderabad along with the Ministry of New and Renewable Energy.

5. Analysis of the seven geothermal provinces of India

Data from over 500 geothermal sites across the seven Indian

provinces was collected for analysis from the official site of Ministry of New and Renewable Energy, Govt. of India. Geothermal Atlas of India [11] and Geothermal Resources of India[12] document the data from the geothermal hotspots spanning all over the country. It contains data for the chemical composition and concentration of various ions for every site. TDS, Silica concentration, Temperature and pH which form the basis of primary indication of scaling, were studied and general trends governing their interdependencies were found out. Data was segregated into Himalayan Province and other six provinces were clubbed into non-Himalayan provinces for better analysis and comparison. Himalayan Province was found to be roughly on the higher side on all of these parameters in comparison to other non-Himalayan provinces. Himalayan province includes all the regions lying along the Great Himalayan belt including the areas Jammu and Kashmir, Uttarakhand, Himachal Pradesh and Northeastern hilly areas.

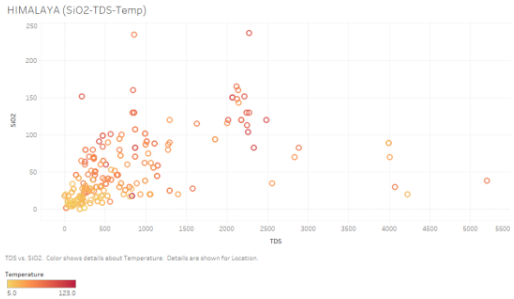


Fig. 2. Concentration of SiO_2 vs. TDS curve with temperature as color gradient from sites of Himalayan Province

As the plot for Himalayan Province sites between concentration of SiO_2 and TDS (Fig. 2.) shows there are more dots on the red side of the temperature scale which indicates how temperature close and above 100°C are common in the region which can be used for the heating purposes as well power generation. Sites like Tapovan, Yurdu, Jamnotri and Manikaran were found to have temperatures on the higher side with comparatively lesser concentration of TDS and SiO_2 which makes them ideal for plant setup for power generation. Tapoban is already under the lens for the scope of power plant setup by the government agencies. On the other hand sites from Puga Valley, Ladakh were found to have the temperatures on the higher side of spectrum but with also a significant presence of Silica and TDS. This means this site will be prone to scaling and would require regular chemical treatment or silica extraction. Puga Valley is an important site for Geothermal in India which has been the centre of focus and poses high potential for meeting the targets of Ministry of New and Renewable Energy (MNRE) of power generation through Geothermal.

The plot between TDS and Temperature from the sites of Himalayan Province (Fig. 3) shows more and more dots with blue color which means more sites have pH more than 7.5 which is when the scaling occurs. Going by just the pH cut off mark, it makes most of the Himalayan sites prone to scaling. The general trend from the data of the sites also suggests how TDS increases gradually and exponentially with the increase in temperature.

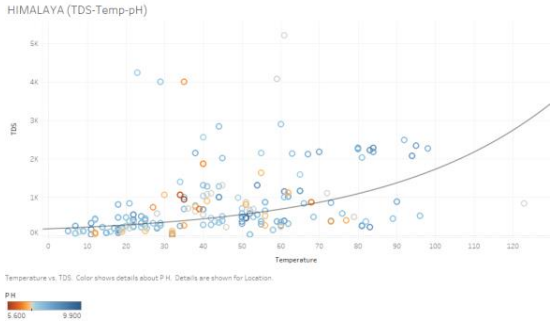


Fig. 3. TDS vs Temperature curve with pH as color gradient from sites of Himalayan Province

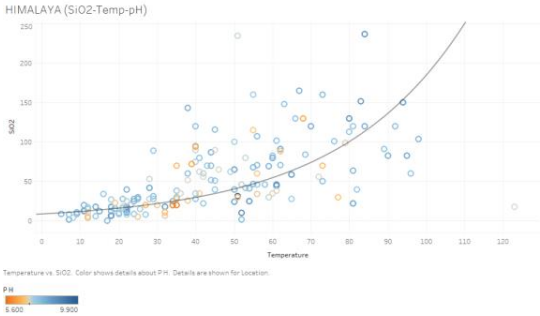


Fig. 4. Concentration of SiO_2 vs. Temperature curve with pH depicted as color gradient from sites of Himalayan Province

The plot between concentration of SiO_2 and Temperature (Fig. 4) also suggests that there is exponential increase in the concentration of Silica as the temperature of the water increases. This indicates that the sites which will be targeted for power generation with higher temperatures will be prone to siliceous scaling and have concentrations in extractable ranges for Silica.

The plot between the concentration of SiO_2 and TDS for the six non Himalayan provinces (Fig. 5) is found to be scattered one but a major cluster for almost all the six provinces is found to be located on lower TDS and lower silica concentration sides are found. This points to lower scaling and corrosion problem in these areas. This can also be attributed to the moderate temperatures of geothermal water in these locations. These moderate temperature ranges are useful for domestic purposes and industries where high temperatures are not needed. Since scaling tendencies are low in these regions overall, there won't be any need for extra care and expenditure to prevent the same.

The plot of TDS and Temperature curve for the six non Himalayan provinces (Fig. 6) depicts the general trend of exponentially increasing TDS with increase in temperature for four out of the six provinces. It comes out to be gradually and slightly increasing for Godavari Province and almost a constant

trend for Mahanadi Province. Mahanadi remains a low TDS zone for all the temperature ranges and all the geothermal sites.

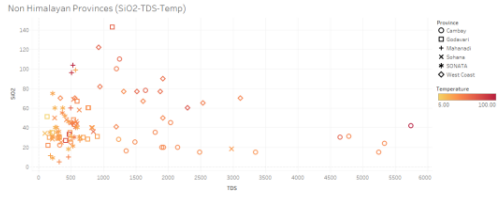


Fig. 5. Concentration of SiO_2 vs. TDS curve with Temperature as color gradient from sites of all six non Himalayan Provinces

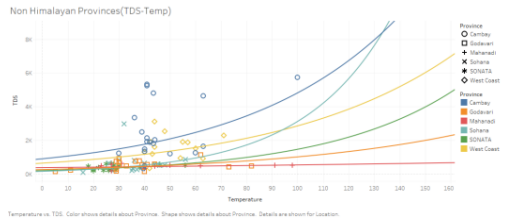


Fig. 6. TDS vs. Temperature curve from sites of all six non himalayan provinces

Tattapani sites also rank high on the temperature scale which makes them the best prospect for power generation as well as industrial heating usage. There is manageable risk of scaling with sufficiently high temperature. This also goes in line with the government's efforts to harness Tattapani geothermal resource alongside Puga Valleya and Tapoban sites in the Himalayan region.

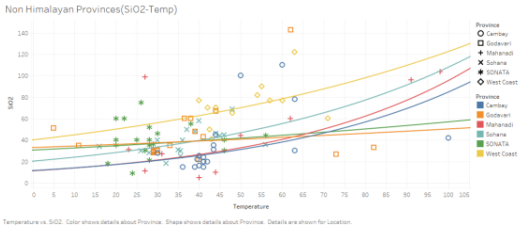


Fig. 7. Concentration of SiO_2 vs. Temperature curve from sites of all six non Himalayan provinces

The plot between concentration of SiO_2 vs. Temperature (Fig. 7) does not reveal any general trend for these six provinces. However, it appears to be scattered with some clusters of individual provinces in the middle which can be due to the compact and small areas of these provinces. Due to this there is some scope of uniformity in the temperature ranges and chemical composition. The trend lines depict increasing tendency of the silica concentration with the temperature.

6. Conclusion

Study and analysis of the data of indicators for scaling and corrosion such as pH, Temperature, concentrations of SiO_2 and TDS was carried out. It was observed that the Himalayan provinces appeared to have higher temperatures in general than the non-Himalayan provinces sites. There is exponential increase in TDS and silica concentration with increase in Temperature. Puga Valley in Ladakh, Tattapani in Chhattisgarh and Tapoban in Uttarakhand were found to be ideal sites for plant setup in line with the government agencies' efforts in tapping geothermal resources at these sites. Mahanadi, Godavari, SONATA provinces were found to have lower TDS values in general which make them viable sites for domestic and moderate industrial usage without high hazards of scaling and corrosion.

References

[1] Mukul Chandra Bora, "Geothermal Energy: Indian Scenario," in National Seminar on Renewable Energy Technologies: Issues & Prospects, At NERIST, Itanagar, India, 2010.
[2] D. Cinti, L. Pizzino, N. Voltattorni, F. Quattrocchi and V. Walia, "Geochemistry of thermal waters along the fault segments in the Beas and Parvati valleys (north-west Himalaya, Himachal Pradesh) and in the Sohna town (Haryana), India," Geochemical Journal, vol. 43, pp. 65-76, 2009.
[3] M. Cabrini, S. Lorenzi, P. Marcassoli, T. Pastore, B. Tarquini, "Corrosion Rate Assessment in Geothermal Water Conveying System," in 2nd ICCMP, Prague, Czech Republic, 2010.
[4] Kevin Rafferty, "Scaling in Geothermal Heat Pump Systems" for U. S. Deptt. Of Energy, Idaho, USA.
[5] Ungemach, Pierre, "Handling of Corrosion and Scaling Shortcomings in Low Enthalpy Geothermal Environments," 2019.
[6] "Chemistry of Thermal Fluids," in Geothermal Systems and Technologies, Geothermal Communities, pp. 36–47.
[7] Einar Gunnlaugsson, "Importance of Chemistry in Geothermal Exploration and Utilization" in Workshop for Decision Makers on Direct Heating Use of Geothermal Resource in Asia, Tianjin, China, May, 2008.
[8] Nikos Andritsos, Pierre Ungemach Petros Koutsoukos, "Corrosion and Scaling," (2014), pp. 181-194.
[9] Burkni Palsson, "Silica and scaling problems in geothermal power plants", openaccessgovernment.org, November, 2014.
[10] Einar Gunnlaugsson, Halldór Ármannsson, Sverrir Thorhallsson and Benedikt Steingrímsson, "Problems in Geothermal Operation – Scaling and Corrosion," in Short Course VI on Utilization of Low- and Medium-Enthalpy Geothermal Resources and Financial Aspects of Utilization, El Salvador, 2014.
[11] "Geothermal Atlas of India", Geological Survey of India, 1991.
[12] Jawahar Lal Thussu, "Geothermal Energy Resources of India", Geological Survey of India, Special Publication-69, 2002.

Applied Water Science (2019) 9:139
https://doi.org/10.1007/s13201-019-1011-z

ORIGINAL ARTICLE

Seasonal behaviour of thermal stratification and trophic status in a sub-tropical Palustrine water body

Priyanka Singh[1] · Jyoti Bagrania[1] · A. K. Haritash[1]

Received: 27 December 2018 / Accepted: 24 June 2019
© The Author(s) 2019

Abstract

Thermal and chemical stratification of water bodies has a strong bearing upon the water quality since it regulates mixing of nutrients/pollutants and affects the aquatic life in it. The present study was undertaken in a shallow pond ecosystem to study thermal stratification and trophic status over a period of 4 months (January–April, 2015) to understand the seasonal variations associated with. Chemical stratification with respect to dissolved oxygen was also observed along the depth. The pond was weakly to strongly thermally stratified throughout the study. The redox potential of water varied from − 246 mV (February 15) to 137 mV (March 15) indicating dominantly reducing conditions. The dissolved oxygen profile with depth confirmed that reducing conditions prevailed in the hypolimnion. The surface epilimnion represented higher values of Trophic State Index for total phosphate (TSI-TP) and relatively lower values of total chlorophyll (TSI-Chl). The value of buoyancy frequency (N) and Richardson number (R) indicated complete and strong mixing in surface epilimnion due to the temperature, wind velocity, and current speed during the study with relatively stronger stratification and lower surface mixing during early summer (March–April). The higher value of TSI-TP compared to TSI-Chl indicated that zooplankton grazing or limitation of nitrogen plays a role in limiting eutrophication and algal growth. Although the wind velocity may be higher during summer, stronger thermal stratification can be formed due to temperature difference between upper layer and lower layer resulting in lower mixing of surface epilimnion, which may be a cause of concern if dissolved phosphate levels are high.

Keywords Trophic State Index · Thermal stratification · Eutrophication · Water quality · Buoyancy frequency (N) · Richardson number (R)

Introduction

Cultural eutrophication is a process that speeds up the rate of natural eutrophication by increasing the inflow rate of nutrients (nitrogen and phosphorous) into the water body. Enrichment of nutrients can lead hyper-eutrophication. Similarly, turbidity of the water body is an important parameter to determine the water quality, and most of the eutrophicated lakes have low transparency because of algal growth, and dead and decomposing organic matter. It results in reduced sunlight penetration, reduced photosynthesis, and low productivity of the lake (Haritash et al. 2015a, b). This process may result in oxygen depletion in the water body as a response to reduced photosynthesis and increased level of nutrients (Balakrishnan et al. 2017). The algal blooms have a number of adverse effects on water users, because algae can reduce dissolved oxygen, affect nutrient cycle produce toxins, and decrease aesthetic water quality (Brown et al. 1998; Haritash et al. 2016). Lakes are generally classified as oligotrophic, mesotrophic, eutrophic, and hypertrophic based on the trophic status. Trophic state is defined as the total weight of the biomass in a water body at a specific location and time (Devi Prasad and Siddaraju 2012). Carlson (1977) used algal biomass as key description for measuring the Trophic State Index because algal blooms are of concern to the public. Many physical factors are responsible of the trophic state of water body; thermal stratification is one of them where density difference in the lake is primarily due to

✉ A. K. Haritash
akharitash@dce.ac.in

Priyanka Singh
singhpriyanka8285@gmail.com

Jyoti Bagrania
bagrania94@gmail.com

[1] Department of Environmental Engineering, Delhi Technological University, Bawana Road, Shahbad Daulatpur, Rohini, Delhi 110042, India

Published online: 08 July 2019

the temperature variation (Yang et al. 2018). Thermal stratification leads to reduction in turbulent mixing and accumulation of algal population in the surface layer. Blooms do not occur under complete mixing of the water column (Bormans et al. 1997). Permanent stratification has decisive impact on redistribution of dissolved substances, in deep lakes complete mixing rarely occurs, and it leads to anoxic condition at bottom of the lake (Boehrer and Schultze 2008; Mellios et al. 2015). A lake is classified as stratified if the temperature difference of 0.2 °C/m or higher with depth is present near the surface of the water body (Pernica and Wells 2012). Thermal stratification is affected by factors, such as solar radiation, turbidity, discharge, and wind speed; and to lower extent to humidity and air temperature. Wind-driven mixing in epilimnion of deep lakes has been studied by Pernica and Wells (2012), and a positive correlation of wind-induced mixing with thermal stratification using gradient Richardson number is mentioned. In this study, thermal stratification is correlated with wind-induced mixing and trophic state to see the effect of thermal stratification on eutrophication.

Materials and methods

This study was carried out from January 15 to April 15 in a shallow pond located in the campus (28°44′58.55″N, 77° 0.6′46.75″E) of Delhi Technological University, Delhi. The pond has total area of 9468 m^2 and average depth of 1.8 m. It receives wastewater from neighbouring residential colony and partially from a university hostel.

In order to achieve average water quality, the pond was divided into 09 grids, and a sample was collected from each grid to evaluate the spatial variations, if any. The samples were collected in pre-rinsed PET bottles and were immediately transported to laboratory (within 10 min) for its characterisation. The samples were characterised for pH, oxidation–reduction potential (ORP), temperature, dissolved oxygen (DO), available phosphate (AP), total phosphate (TP), nitrogen (as nitrate and total Kjeldahl nitrogen TKN), calcium, magnesium, chlorophyll (Chl), Secchi depth (SD) as per standard methods in triplicates using analytical grade (AR) chemicals and ultrapure (type I) water. The parameters like pH, temperature (°C), DO (mg/l), ORP (mV) were measured on the site using Orion make Star A329 model multiparameter meter, and SD (m) and total depth (meter) were also measured on site using Secchi disc and a graduated steel rod, respectively. TKN (mg/l) was measured using Kjeldahl nitrogen estimation system. Extraction and determination of chlorophyll were performed according to the method prescribed by Arnon (1949). Other parameters were determined within 24 h using standard methods as prescribed by APHA (1995).

Trophic status

A popular method for examining trophic state of a water body uses the Trophic State Index (TSI) developed by Carlson (1977) to determine Carlson Trophic State Index (CTSI). Chlorophyll was calculated by using following equation given by Arnon (1949).

$$\text{Chlorophyll 'a' (mg/ml)} = (0.0127) \times (A663) - (0.00269) \times (A645) \quad (1)$$

where A663 and A645 are the absorbance of extract at 663 nm and 645 nm wavelength using spectrophotometer (LabIndia make UV3092 model). The formulae for calculating the TSI values where SD is Secchi depth (m), Chl is chlorophyll (mg/m^3), and TP is total phosphorus (mg/l) are as follows:

$$\text{Secchi disk:} \quad \text{TSI(SD)} = 60 - 14.41 \ln (\text{SD}) \quad (2)$$

$$\text{Chlorophyll a:} \quad \text{TSI(CHL)} = 9.81 \ln (\text{Chl}) + 30.6 \quad (3)$$

$$\text{Total phosphorus:} \quad \text{TSI(TP)} = 14.42 \ln (\text{TP}) + 4.15 \quad (4)$$

The following equation can be used to compute the CTSI:

$$\text{CTSI} = \{(\text{TSI} - \text{SD}) + (\text{TSI} - \text{Chl}) + (\text{TSI} - \text{TP})\}/3 \quad (5)$$

Wind

Data of wind speed (u) of the study area were taken from the Indian Meteorological Department (IMD) for the months of January 15 to April 15. Wind-induced current was empirically calculated, for wind speeds between 100 and 500 cm/s "the wind factor (current speed/wind speed) decreased linearly with wind speed". At wind speeds above 500 cm/s, "the wind factor remained in the range of 1–3%" (George 1981). For the present study, the wind factor was considered as 2% of wind speed (cm/s). The occurrence of stable stratification depends on the relative strengths of turbulent mixing and thermally induced density gradients (Richardson 1926). The relative magnitudes of these two processes have been calculated by the non-dimensional ratio known as the Richardson number, R, expressed as:

$$R = N^2/(\mathrm{d}u/\mathrm{d}z)^2 \quad (6)$$

where buoyancy frequency N is define as $N = ((g/\varrho)\ (\mathrm{d}\varrho/\mathrm{d}z))^{1/2}$, g = acceleration due to gravity (9.81 m/s^2), ϱ = density of fluid (water, kg/m^3), u = velocity (m/s) and z=depth (m) (Pernica and Wells 2012). Denominator of Eq. (6) is the vertical velocity shear. According to (Pernica and Wells 2012), there will be little or no turbulent mixing for $R < 0.25$.

Results and discussion

Based on the results obtained, the shallow pond in DTU receives significantly high concentration of dissolved phosphate and it undergoes thermal stratification seasonally. The water body was classified as eutrophic throughout the study, and the details of water quality, thermal stratification, and the trophic status are discussed in the following sections.

Water quality

The water quality on the basis of physico-chemical characterisation of collected water samples revealed that the water lies in slightly alkaline range with average pH ranging from 7.6 to 8.7 (Table 1) with dominantly reducing conditions based on values of oxidation–reduction potential (ORP). The average ORP varied from −246 (February 15) to 137 mV (March 15). The concentration of available and total phosphate was observed to be significantly high considering its potential (more than 1.0 mg/l) for eutrophication. Phosphorous which is generally recognized as the key nutrient for algal growth is an essential element to determine the fertility of ponds (Shah and Pandit 2012). The concentration of available phosphate (orthophosphate phosphorus) during the study period ranged from a maximum of 9.4 ± 1.2 mg/L under reducing condition (ORP −299 mV) in February to a minimum of 7.1 ± 0.9 mg/l under oxidising condition (ORP 137 mV) in March. The concentration of dissolved phosphate remains lower under oxidising conditions since binding of PO_4^{3-} takes place with Fe^{3+} and Al^{3+} ions which forms an insoluble metal–phosphate complex under oxidising conditions, whereas under reducing conditions Fe^{3+} ions are reduced to Fe^{2+} which breaks the metal–phosphate complex resulting in remobilisation of phosphate in water (Haritash et al. 2017). Increase in the orthophosphate concentration during highly reducing conditions has also been observed by Olila and Reddy (1997). In the present study, phosphate concentration was significantly high compared to the other nutrient, i.e. nitrogen expressed as nitrate (mg/l) and TKN (mg/l); most of the nitrogen was present as ammoniacal or organic (as TKN) in the form relatively lower concentration of nitrate (from 1.1 to 6.9 mg/l) compared to ammoniacal nitrogen further strengthens the observation that the condition in the pond were chiefly reducing during the study.

Trophic status of pond

The trophic state indices for phosphate (TSI-TP), chlorophyll (TSI-Chl), and Secchi depth (TSI-SD) are used to arrive at an aggregated average referred to as Carlson's Trophic State Index (CTSI). The TSI values of different parameters are given in Table 2. It is observed that TSI-TP has maximum value followed by TSI-Chl and TSI-SD. The combined CTSI was observed to more than 100 (Table 3) in all the cases indicating that the pond is eutrophic throughout the study; CTSI value from the observation was found to be maximum in April (117) and minimum in March (113) due to low concentration of phosphate in water during March. Since TSI-TP > TSI-Chl > TSI-SD, the transparency of water is not affected significantly limiting the effects of eutrophication. This may also be attributed to zooplankton grazing and/or limitation of nitrogen. In this study, both the factors were observed to play a significant role in keeping TSI-SD values less than the other parameters (Fig. 1).

Stratification of pond

The observed values of temperature and dissolved oxygen with respect to depth revealed that the pond remained stratified (weak to strong) during the study (Fig. 2). The temperature lapse was observed to be 0.8, 1.1, 4.4, and 3.0 during months January, February, March, and April of 2015 (Table 1), respectively. As per the literature available,

Table 2 TSI values for SD, TP, Chl(a) in pond during the study

TSI/months	Jan 15	Feb' 15	Mar 15	Apr 15
TSI(SD)	80	82	79	82
TSI (Chl)	127	128	124	133
TSI(TP)	135	137	136	137
CTSI	114	116	113	117

Table 1 Monthly variations of physico-chemical characteristics, thermal stratification (d*T*/d*z*), buoyancy frequency (*N*), Richardson number (*R*) during the study

Parameters/months	pH	ORP (mV)	Temp. (°C)	AP (mg/l)	TP (mg/l)	Nitrogen (mg/l)		d*T*/d*z* (°C/m)	*N* (s^{-1})	*R*
						NO_3^-	TKN			
Jan 15	8.1 ± 0.1	108 ± 48	14.2 ± 0.5	8.4 ± 0.3	9.3 ± 2	3.2 ± 0.5	13.0 ± 4.5	0.8	0.031	2.18
Feb 15	7.6 ± 0.1	−246 ± 32	15.3 ± 0.1	10.3 ± 2.3	13.7 ± 1.9	1.1 ± 1.0	16.3 ± 1.2	1.1	0.039	1.37
Mar 15	8.7 ± 0.2	137 ± 26	21.1 ± 1.3	9.3 ± 0.9	11.6 ± 1.7	6.9 ± 0.8	9.5 ± 2.1	4.4	0.101	1.91
Apr 15	8.3 ± 0.1	120 ± 74	26.3 ± 0.9	8.6 ± 1.7	14.3 ± 1.9	2.0 ± 0.4	10.3 ± 0.5	3.0	0.085	1.02

Table 3 Carlson's Trophic State Index values and classification of lakes (Devi Prasad and Siddaraju 2012)

TSI values	Trophic status	Attributes
< 30	Oligotrophic	Clear water, oxygen throughout the year in the hypolimnion
30–40	Oligotrophic	A lake will still exhibit oligotrophy, but some shallower lakes will become anoxic during the summer
40–50	Mesotrophic	Water moderately clear, but increasing probability of anoxia during the summer
56–60	Eutrophic	Eutrophic lower boundary of classical eutrophy: decreased transparency, warm-water fisheries only
60–70	Eutrophic	Eutrophic dominance of blue–green algae, algal scum probable, extensive macrophyte problems
70–80	Eutrophic	Eutrophic heavy algal blooms possible throughout the summer, often hypereutrophic
> 80	Hypereutrophic	Eutrophic algal scum, summer fish kills, few macrophytes

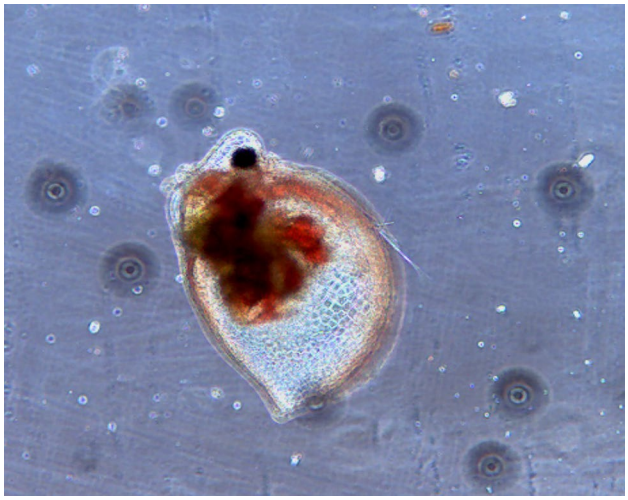


Fig. 1 Microscopic view (100×) of *Ceriodaphnia* spp. present in the DTU pond

a water body is classified as thermally stratified if dT/dz is greater than 0.1 °C/m. Based on the values of thermal stratification, it was observed that the pond was weakly stratified during winter (January and February) as compared to the summer (March and April). Physical and chemical both forces are responsible for the process of stratification. Temperature and dissolved substances contribute to density difference in water (Boehrer and Schultze 2008; Mellios et al. 2015). Variation of temperature with depth is described in Fig. 2. A sharp change of temperature from surface to bottom was observed during summer with the temperature difference of 8.1 °C in March and 6.5 °C in April; as against the difference of 1.4 °C in January and 2.1 °C in February. If temperature of surface water falls below the temperature of the underlying layer, the entire water body can overturn (Boehrer et al. 2008). Starting from homogeneous conditions in winter, surface water is heated by radiation and contact with the atmosphere, and mixing usually driven by wind due to which, overturning was not found in the pond. Thermocline was found at approximately same depth of 3 °cm during the study.

Wind-induced stratification

Temperature, wind velocity, current speed and dissolved oxygen data are compared in this section, and observations are described in Table 1 and Fig. 2. Anoxic hypolimnion was observed during the study. Further analysis of this data, including calculation of Buoyancy frequency and Richardson number (R), is also included. Buoyancy frequency and Richardson number having its critical value $N_{cri} = 0.025\ s^{-1}$, $R_{cri} = 0.25$ (Pernica and Wells 2012) below which little or no mixing is observed, ranged from maximum of 2.18 during January to a minimum of 1.02 in April (Table 1). Near-surface stratification was found to be weak during April with minimum R value, because epilimnion gets completely mixed due to the high wind speed. Strong winds (above about 3 m/s) induce Langmuir circulations (large-scale counter-rotating helical vortices) which is largely responsible for mixing of the surface layer and deepening of the epilimnion, but an average wind speed (1–2 m/s) can cause a complete mixing of stratified surface layer of the water body (Chubarenko et al. 2010).

Conclusion

Based on the physico-chemical characteristics, the water body was observed to be rich in phosphate and deficient of nitrogen. Although the Trophic State Index (TSI) due to phosphate was high, its CTSI values were lower than TSI-TP, indicating that eutrophication in the ecosystem is regulated by limited nitrogen and zooplanktons grazing. Based on temperature profile with depth, the pond was classified as stratified with temperature lapse more than 0.1 °C/m throughout the study. Significant difference between surface and bottom temperature confirmed the thermal stratification. Moreover, chemical stratification in terms of DO concentration was also observed with significant low concentration of DO at bottom within the average depth of 1.8 m, indicating the strong stratification may have severe impact on bottom-dwelling aerobic organisms. The prevalence of reducing conditions in bottom layer resulted in solubilisation/mobilisation of phosphate

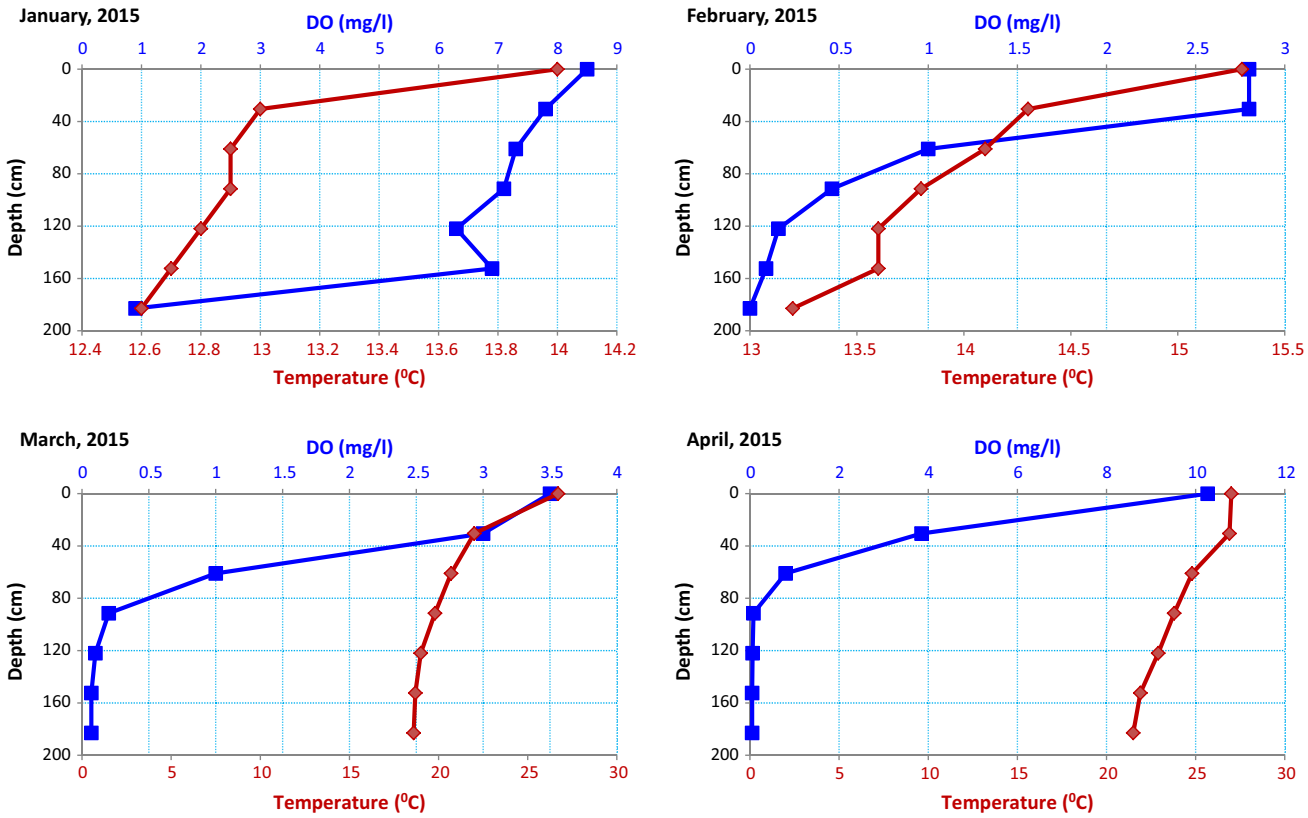


Fig. 2 Thermal (temperature—°C) and chemical (DO—mg/l) stratification in DTU pond during the study

in bottom layer. This may result in excessive algal growth upon overturning of the lake (when the thermal stratification breaks). Therefore, it becomes important to adopt effective interventions which result in reduction in import of phosphate and enhance its export, e.g. chemical precipitation of phosphate followed by sediments dredging and/or promoting aquaculture and fish catch. Managing water quality even of shallow ponds is ecologically important to maintain its water quality especially in respect of eutrophication and sustain natural ecosystem within it.

Acknowledgements The authors acknowledge the help of DTU administration for providing necessary facilities to undertake the study.

Compliance with ethical standards

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

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

APHA (1995) Standard methods for the examination of water and waste water. AWWA, WPCE, Washington

Arnon DI (1949) Copper enzymes in isolated chloroplasts, polyphenoxidase in beta vulgaris. Plant Physiol 24:1–15

Balakrishnan S, Chelladurai G, Mohanraj J, Poongodi J (2017) Seasonal variations in physico-chemical characteristics of Tuticorin coastal waters, southeast coast of India. Appl Water Sci 7:1881–1886

Boehrer B, Schultze M (2008) Stratification of lakes. Am Geophys Union 46:1–27

Boehrer B, Fukuyama R, Chikita K (2008) Stratification of very deep, thermally stratified lakes. Geophys Res Lett 35:1–5

Bormans M, Maier H, Burch M, Baker P (1997) Temperature stratification in the lower River Murray, Australia: implication for cyanobacterial bloom development. Mar Freshwater Res 48:647–654

Brown CD, Canfield DE, Bachmann RW, Hoyer MV (1998) Seasonal patterns of chlorophyll, nutrient concentrations and Secchi disk transparency in Florida lakes. J Lake Reserv Manag 14(1):60–76

Carlson RE (1977) A trophic state index for lakes. Limnol Oceanogr 22(2):361–369

Chubarenko I, Chubarenko B, Esiukova E, Baudler H (2010) Mixing by Langmuir circulation in shallow lagoons. Baltica 23:13–24

Devi Prasad AG, Siddaraju P (2012) Carlson's Trophic State Index for the assessment of trophic status of two lakes in Mandya district. Adv Appl Sci Res 3(5):2992–2996

George DG (1981) Wind-induced water movements in the South Basin of Windermere. Freshw Biol 11:37–60

Haritash AK, Shan V, Singh P, Singh SK (2015a) Preliminary investigation of environmental status of Bhindawas Bird Sanctuary. Int J Eng Res Technol 4(3):53–56

Haritash AK, Yadav A, Kataria A, Singh K, Mathur K, Goswami S (2015b) Seasonal assessment of trophic state of a Palustrine water body. Int J Eng Res Technol 4:37–40

Haritash AK, Gaur S, Garg S (2016) Assessment of water quality and suitability analysis of River Ganga in Rishikesh, India. Appl Water Sci 6:383–392

Haritash AK, Dutta S, Sharma A (2017) Phosphate uptake and translocation in a tropical canna-based constructed wetland. Ecol Process 6:12. https://doi.org/10.1186/s13717-017-0079-3

Mellios N, Kofinas D, Laspidou C, Papadimitriou T (2015) Mathematical modeling of trophic state and nutrient flows of lake Karla using the PCLake model. Environ Process 2(1):S85–S100

Olila OG, Reddy KR (1997) Influence of redox potential on phosphate uptake by sediments in two sub-tropical eutrophic lakes. Hydrobiologia 345:45–57

Pernica P, Wells M (2012) Frequency of episodic stratification in the near surface of lake Opeongo and other small lakes. Water Qual Res J Can 47(3–4):227–237

Richardson LF (1926) Atmospheric diffusion shown on a distance-neighbour graph. Proc R Soc Ser A 110:709–737

Shah JA, Pandit AK (2012) Physico-chemical characteristics of water in Wular lake—a Ramsar site in Kashmir Himalaya. Int J Geol Earth Environ Sci 2(2):257–265

Yang Y, Wang Y, Zhang Z, Wang W, Ren X, Gao Y, Liu S, Lee X (2018) Diurnal and seasonal variation of thermal stratification and vertical mixing in a shallow fresh water lake. J Meteorol Res 32:219–232

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

Send Orders for Reprints to reprints@benthamscience.net

Current Smart Materials, 2019, *4*, 36-44

RESEARCH ARTICLE

Studies on Polyacrylate-Starch/Polyaniline Conducting Hydrogel

Reetu Prabhakar* and Devendra Kumar

Department of Applied Chemsitry & Polymer Technology, Delhi Technological University, New Delhi-110042, India

Abstract: ***Background*:** The superabsorbent polymers based conducting hydrogel such as polyaniline impregnated polyacrylate-starch hydrogel were synthesized *via* two -steps interpenetrating polymer network process. In the present work instead of using a synthetic polymer of acrylamide, a biodegradable polymer such as starch has been used with polyacrylate superabsorbent polymer. The main attempt of this work is to analyze the electrical conductivity of resulting hydrogel at varying concentrations of crosslinker, initiator, monomer, and a copolymer for improving the properties of synthesized hydrogel and elaborating the diversity of its utilization.

***Methods*:** The polymerization of aniline was performed through the absorption of aniline monomer into the polymer matrix followed by the addition of initiator/dopant solution. The morphological and structural analysis and thermal stability of the synthesized hydrogel were studied using Scanning Electron Microscopy (SEM), Fourier transform infrared spectroscopy and thermogravimetric analysis (TGA), respectively. The swelling behaviour of the synthesized hydrogel was performed in a different medium.

***Results*:** Electrical conductivity data of polyacrylate-starch/polyaniline (PAANa-starch/PANI) were compared from polyacrylate-co-polyacrylamide/polyaniline P(AANa-co-AM)/PANI hydrogel, which revealed that polyacrylate-starch/polyaniline shows higher conductivity than polyacrylate-co-polyacrylamide/polyaniline conducting hydrogel. Maximum swelling of the synthesized hydrogel was reported in the basic medium.

***Conclusion*:** It is observed that the addition of starch into the matrix significantly improved the overall properties of hydrogel. The polymerization of aniline was done *in-situ* with the absorption of aniline monomer into the PAANa-starch polymer matrix followed by soaking in an initiator/dopant solution. The XRD pattern of PAANa-starch/PANI showed a broad peak at 22.8° while no peak was observed in the PAANa-starch gel, implying that PAANa-starch/PANI has a crystalline and more ordered structure. PAANa-starch/PANI has higher conductivity than the P(AANa-co-AM)/PANI hydrogel. This enhanced electrical conductivity in case of PAANa-starch/PANI hydrogel could be due to the more crosslink points of synthetic polymer polyacrylamide between PAANa-PANI hydrogel.

ARTICLE HISTORY

Received: April 15, 2018
Revised: February 27, 2019
Accepted: March 06, 2019

DOI:
10.2174/2405465804666190313154427

Keywords: Electrical conductivity, polyacrylate, surface morphology, hydrogel, conducting polymers, superabsorbent polymers.

1. INTRODUCTION

Superabsorbent hydrogels based on superabsorbent polymers, such as polyacrylic acid, polyacrylamide and neutralized polyacrylic acid or polyacrylate are three dimensional polymer network structures that have the capability of absorbing water up to hundred to thousand times. These swelled hydrogels not only absorb water to form stable colloid but also hold the same under applied pressure [1-3].

In recent years, there is a progress in hydrogels based on superabsorbent polymers with Conducting Polymers (CP) for the preparation of multifunctional electrical conducting materials. Conducting hydrogels find applications in various fields including biological, optoelectronics, dye sensitized solar cells and rechargeable lithium ion batteries [4-8]. The resulting hydrogel exhibits superior properties such as good processability, chemical stability towards dopants and solubility under readily accessible conditions [5]. In spite of their distinct characteristics, the poor conductivity of hydrogels has limited its practical applications. Hence, much attention has been paid by the researchers towards incorporating more conducting materials into hydrophilic polymer matrix to form a conducting hydrogel. Several successful efforts have been made to modify the hydrogels with different conducting material such as organic conjugated polymers like polyaniline, polypyrrole or inorganic material such as Cu, Zn,

*Address correspondence to this author at the Department of Applied Chemsitry & Polymer Technology, Delhi Technological University, New Delhi-110042, India; Tel: +91 9891694105;
E-mail: drreetu1985@gmail.com

2405-4666/19 $58.00+.00 © 2019 Bentham Science Publishers

Current Smart Materials

graphite, *etc*. Several successful attempts have been made to investigate the effects of the synthesis conditions, such as concentration of the initiator, crosslinker and monomer, reaction temperature, and the ratio of conductive filler on the electrical conductivity of the hydrogel [9-15].

Nowadays, an Interpenetrating Polymerization (IPN) method has attracted much attention worldwide for the synthesis of conducting hydrogel composed of conducting polymer with superabsorbent polymers like polyacrylic acid, polyacrylamide, and neutralized polyacrylic acid or polyacrylate. Among conducting polymers, polyaniline has gained much interest for various applications such as batteries, sensors, conductive fibers and textiles, and fluid separation membranes [16-19] because of its ease of synthesis, good environmental stability, low cost and high conductivity [20, 21].

In the synthesis approach, a superabsorbent hydrogel matrix is immersed in aniline monomer followed by initiator aqueous solution. After *in-situ* polymerization, polyaniline (PANI) will be formed in the matrix which will result in the formation of conducting hydrogel. A single hydrophilic group, such as carboxyl (-COOH) in polyacrylate, amide ($-CONH_2$) in polyacrylamide cannot provide good absorbency. The enhancement of absorbing capacity of hydrogel matrix becomes essential for the formation of conducting hydrogels with enhanced conducting behaviour. In few reports, it has been reported that the mutually absorbent effect of $-CONH_2$, -COONa, -OH, and -COOH groups is better to that of single $-CONH_2$, -COONa, -OH or -COOH groups [22-24]. The introduction of natural polysaccharides has been growing interest due to their biodegradability, biocompatibility, and non-toxicity. Starch is one of the essential polysaccharides that exist mostly in grain/cereals or tubers like potato *etc*. The chemical variation of starch or modified starch composes the most important field for improving the properties of starch and widening the diversity of its utilization.

In this work, a polyacrylate-starch/polyaniline (PAANa-starch)/PANI conducting hydrogel was synthesized *via* an aqueous IPN method at ambient temperature. The polymerization of aniline was done through the absorption of aniline monomer into the polymer matrix followed by addition of initiator/dopant solution. Molecular structure, surface morphology, and crystalline nature were investigated carefully using various characterization techniques. The objective of this work is to analyze the electrical conductivity of resulting hydrogel at a varying concentration of crosslinker, initiator, monomer, and copolymer. Thermal stability and swelling behavior at different medium were also analyzed on PAANa-starch/PANI hydrogel.

2. EXPERIMENTAL

2.1. Materials

Acrylic Acid (AA) monomer and ammonium peroxydisulphate (APS) were purchased from Merck, India. Crosslinker N'N methylene bisacrylamide (N'N,MBA) for the preparation of hydrogel and aniline monomer was purchased from SRL, Pvt. Ltd., India. Aniline monomer was distilled prior to its use. All the solutions used during the experiments were prepared with double distilled water.

2.2. Preparation of (Polyacrylate-Starch) Superabsorbent Polymer

A prerequisite amount of ice-cooled acrylic acid was neutralized by adding sodium hydroxide (NaOH) solution to a preferred neutralization point. Crosslinker N'N,MBA varying from (0.05-0.25 wt%) was gradually added to the mixed solution and degassed under nitrogen atmosphere for 15 min at room temperature to remove oxygen which could slow down the polymerization reaction. In another flask, dried starch was dispersed in distilled water and purged under nitrogen atmosphere. After purging, starch solution was heated at 60°C in a water bath for a few minutes under stirring to obtain the starch slurry. The mixed solution containing neutralized acrylic acid or acrylate, crosslinker N,N'MBA was poured into the starch slurry solution and stirred for few minutes. Under vigorous stirring, an initiator APS varying from 0.05wt% was added to the reaction mixture. The reaction mixture was kept at room temperature to allow the completion of the copolymerization reaction. After gel formation, the product was washed repeatedly two to three times with distilled water to remove unwanted impurities and it was cut into small pieces, dried, milled, sifted and stored away from moisture.

2.3. Preparation of PAANa-starch/PANI Hydrogel

A PAANa-starch/PANI was synthesized according to the following method: firstly PAANa-starch superabsorbent polymer (SAP) was added in an aqueous medium of aniline (0.5-2.75 wt%) under stirring which results in the absorption of aniline monomer inside the SAP matrix and formation of swelled sample. Secondly, swelled sample was then immersed in APS/HCl dopant aqueous solution. The colour of the sample changes slightly from brown to green, which shows the formation of PANI within the SAP matrix. Further, the resulting product was washed with distilled water, dried under vacuum, and then crushed and passed to the sieve to obtain PAANa-starch/PANI powder. PAANa-starch/PANI powder was then dipped in distilled water and kept at room temperature, that results in the formation of PAANa-starch/PANI hydrogel which was filtered and the filtrate was kept aside for further studies.

2.4. Material Characterization

Analytical measurements such as Fourier Transform Infrared (FTIR), UV-Visible and X-Ray Diffraction (XRD) measurements were performed using dried PAANa-starch and PAANa-starch/PANI to investigate their structures. FTIR spectra were recorded using a Thermo Nicolet (Model-380) spectrophotometer with powdered dried gels by the conventional KBr pressed pellets. UV-visible absorption spectra of the hydrogel samples were carried out using a double beam UV-visible spectrophotometer (Cary Series, Agilent Technologies) in the range 200-800 nm. The X-RD was performed on X-ray diffractometer (D8 ADVANCE of Germany Bruker Co.) using Cu Ka radiation source. The surface morphology of dried hydrogel samples was observed

using a Scanning Electron Microscope (SEM) (Hitachi, S-3700N). The fractured hydrogel samples were gold coated before analysis. Thermal analysis was performed using Thermogravimetric Analysis (TGA), TA Instruments, New Castle, DE, USA, (Model Q-50) up to a temperature of 600°C at a heating rate of 10°C/min in nitrogen atmosphere.

2.5. Electrical Conductivity Measurement

The electrical conductivity of hydrogel samples was measured by inserting a pocket conductivity meter (HANNA 8733) in a tube having swollen hydrogel sample. Before the analysis, powdered PAANa/PANI composite was dipped in distilled water at room temperature until equilibrium is reached and filtered over the screen to remove the extra water content.

2.6. Swelling Study

The swelling behavior of crosslinked composite hydrogel was measured in different pH medium. Hydrogel samples were vacuum dried up to a constant weight. After the weight of the dried samples was determined, the samples were equilibrated in swelling media for a day and weighed again. The equilibrium Swelling Ratio (SR) of hydrogel samples were calculated using the following equation:

$$SR = \frac{W_s - W_d}{W_s} \quad (1)$$

where Ws is the weight of the swelled hydrogel in the equilibrium state and W_d is the weight of hydrogel sample in the dry state.

3. RESULTS AND DISCUSSION

3.1. Swelling Behaviour

The equilibrium swelling behaviour of hydrogel was studied in a different medium (Fig. **1**). A maximum swelling was observed in basic medium whereas, lowest occurred in acidic medium as well as in neutral medium. It seems that the two ionic groups $=NH^+$ in acidic and COO^- in basic medium are liable for the sharp change of swelling behavior of hydrogel sample. In acidic condition, a screening effect of the counter ions, *i.e.*, Cl^-, shields the positive charge of $=NH^+$ moiety of aniline and evades an efficient repulsion. As a result, a lowering in equilibrium swelling is observed due to the collapse of hydrogel network structure. However, in neutral medium, acidic and basic functional groups remain in unionized forms, resulting in the formation of hydrogen bonding between amine and carboxylate groups which leads to a decrease in the swelling.

While, in basic medium, carboxylate ions (COO^-) from polyacrylate unit become ionized and undergo electrostatic repulsion between the corresponding COO^- charged sites resulting in maximum swelling potential in hydrogel samples [25-28].

3.2. FTIR Analysis

Fig. (**2**) shows the FTIR spectra of the dried PAANa-starch/PANI (a) and PAANa-starch gel (b). A broad absorption band at 3463 cm^{-1} shown in curve (a) is attributed to O-H and N-H stretching modes in PAANa-starch/PANI hydrogel. Other absorption bands at around 2248, 1718 and 1166 cm^{-1} ascribed as $-CH_2-$ stretching, C=O bending, and C-H bending of polyacrylate unit, respectively. The peaks at about 1616 and 1448 cm^{-1} are responsible for the C-C stretching mode of quinoid and benzenoid ring of PANI [29]. A peak observed at around 815 cm^{-1} is assigned due to the addition of starch with polyacrylate. In the FTIR spectra of PAANa-starch, some shift of peaks has been noticed clearly. The band at around 3440 cm^{-1} is assigned to the N-H stretching band of the PAANa while peaks at 1675 and 1429 cm^{-1} corresponds to the C=O in carboxyl group and C-O in CHCO plane bending which are characteristic peaks for PAANa.

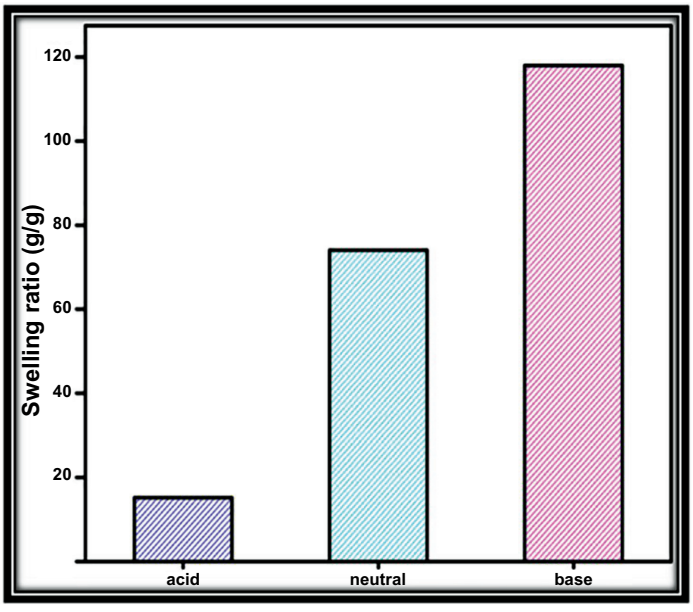


Fig. (1). Swelling behaviour of PAANa-starch/PANI hydrogel in different medium.

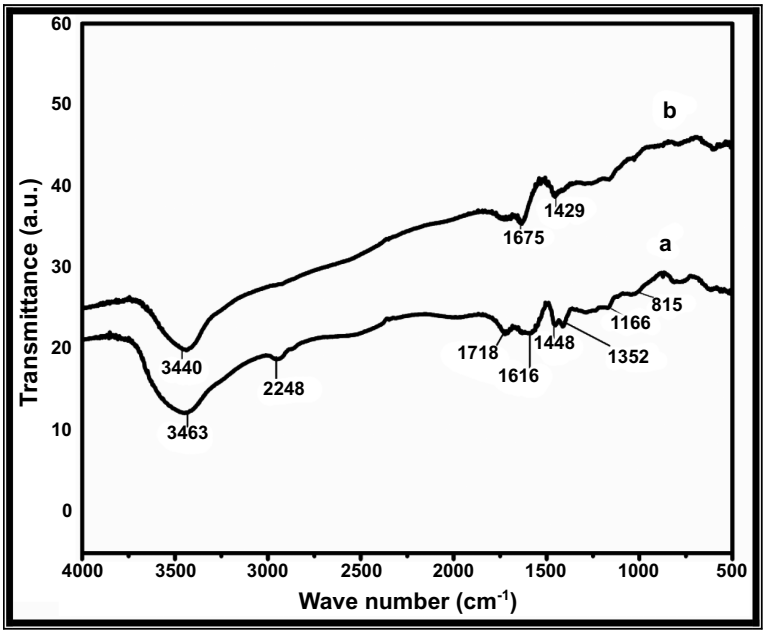


Fig. (2). FTIR spectra of (**a**) PAANa-starch/PANI and (**b**) PAANa-starch hydrogel.

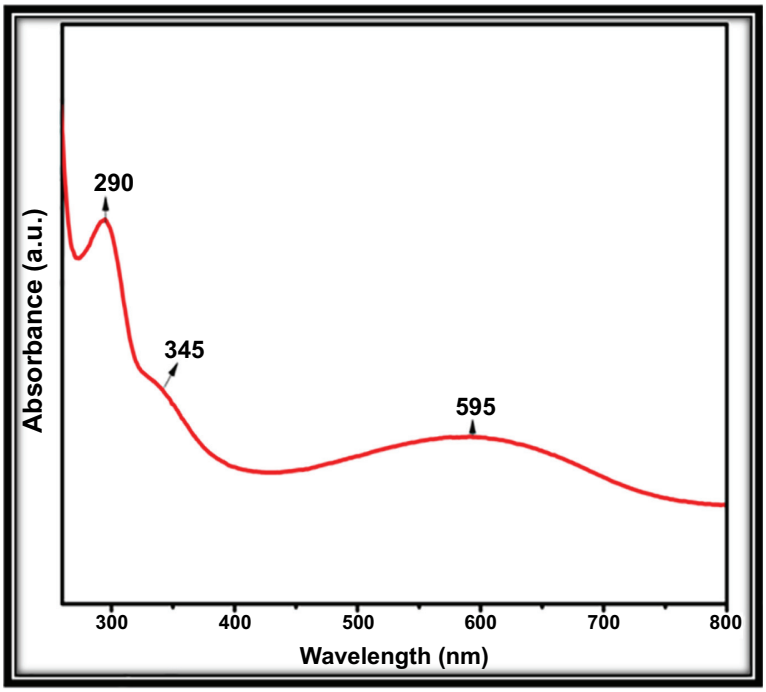


Fig. (3). UV-Visible absorption spectra of PAANa-starch/PANI hydrogel.

3.3. UV-Visible Study

Fig. (**3**) shows the UV-Visible spectra of PANI impregnated hydrogel. The UV-Visible study of PANI impregnated hydrogel was done in the range 280-800 nm. Three characteristic absorption bands appeared in the UV-Vis spectra of PANI embedded hydrogel. The absorption band at 290 and 345 nm are attributed to the π -π^* transition of benzenoid segments and the absorption band at 595 nm is related to the exciton absorption of the quinoid rings [30, 31]. The existence of these absorption peaks indicates the incorporation and formation of emeraldine PANI into the gel.

3.4. XRD Analysis

Fig. (**4**) represents the XRD pattern of the synthesized native PAANa-starch and PANI impregnated PAANa-starch hydrogel samples. The XRD pattern of PAANa-starch/PANI shows a prominent peak at about 22.8° which is the characteristic amorphous peak of PANI [32] while a broad peak appeared at 36.5°, which may be due to the occurrence of PAANa and starch. It implies that PAANa-starch/PANI hydrogel system has crystalline and ordered structure. Thus, the XRD-pattern of the impregnated gel reveals the polymerization and incorporation of PANI within the polymer matrix.

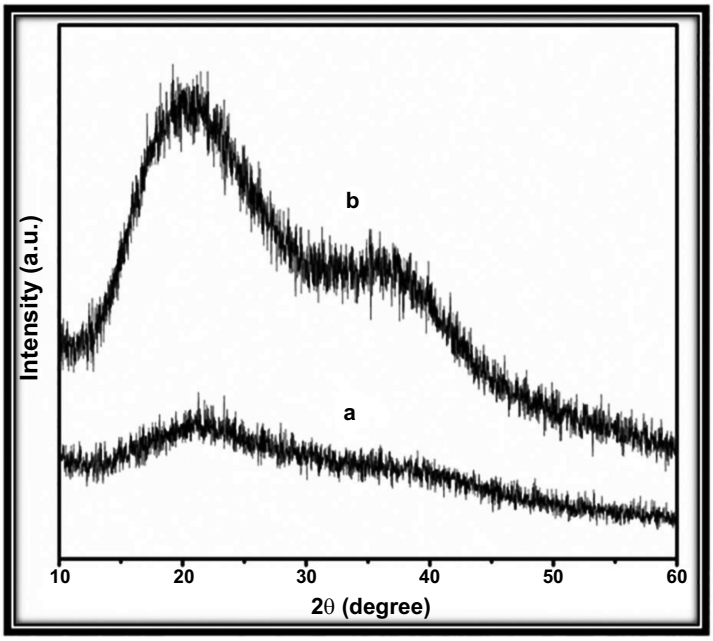


Fig. (4). XRD pattern of (**a**) PAANa-starch and (**b**) PAANA-starch/PANI hydrogel.

The ordered PANI molecular chain is advantageous to the perfection of electric field response of the conducting hydrogel.

3.5. SEM Morphologies of PAANa-Starch and PAANa-Starch/PANI Hydrogels

The SEM micrographs of PAANa-starch and PAANa-starch/PANI hydrogels are shown in Fig. (**5**). As seen in SEM micrographs, the PAANa-starch hydrogel relatively possesses compact layered surface, but PANI impregnated hydrogel form some cluster networks like morphology. It can be implied that the PANI chains polymerized successfully and are well distributed to form network structure within the hydrogel matrix.

3.6. Thermal Analysis

Fig. (**6**) shows the TGA thermograms of PAANa-starch and PAANa-starch/PANI composite hydrogel. A little weight loss at around 100-120°C is observed due to the loss of moisture or water content from the polymer composites. The PAANa-starch thermogram shows a distinct weight loss at 361°C, attributable to the degradation of polymer chains such as PAANa and starch. The composite PAANa-starch/PANI exhibits two steps transitions as seen in Fig. (**6**). The first step examined within the temperature range of 260°C that is associated to the loss of dopant molecules [33] while the weight loss at around 400°C after the removal of the dopant molecules corresponds to the breakdown of the polymer chain [34]. Thus TGA analysis implies that impregnation of PANI improved the thermal stability of composite hydrogel.

3.7. Factors Affecting the Electrical Conductivity of Hydrogel

To get the crosslinked hydrogel with highest conducting ability, the different synthesis ingredients such as initiator,

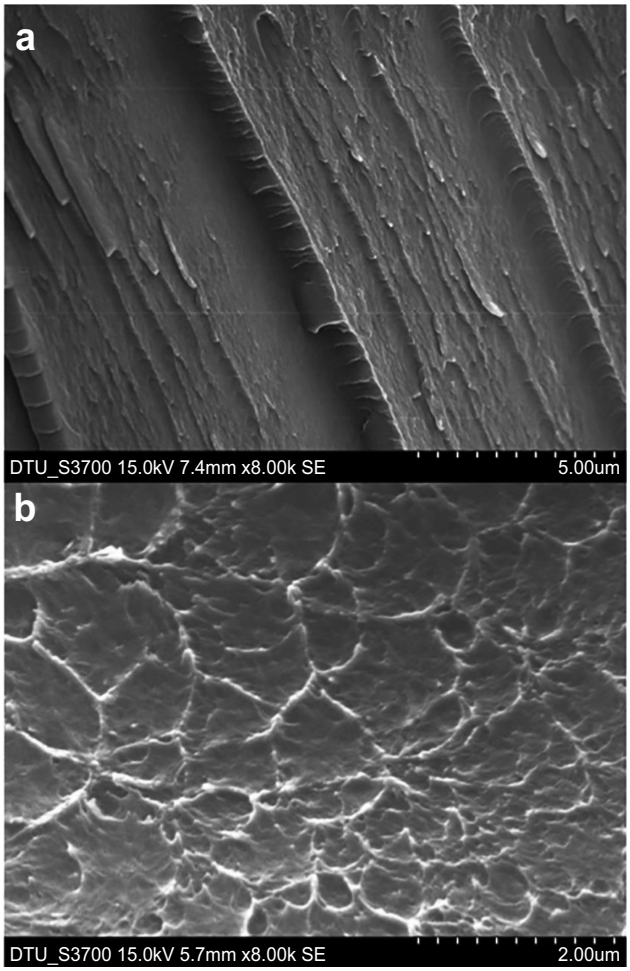


Fig. (5). SEM micrographs of (**a**) PAANa-starch and (**b**) PAANa-starch/PANI hydrogel.

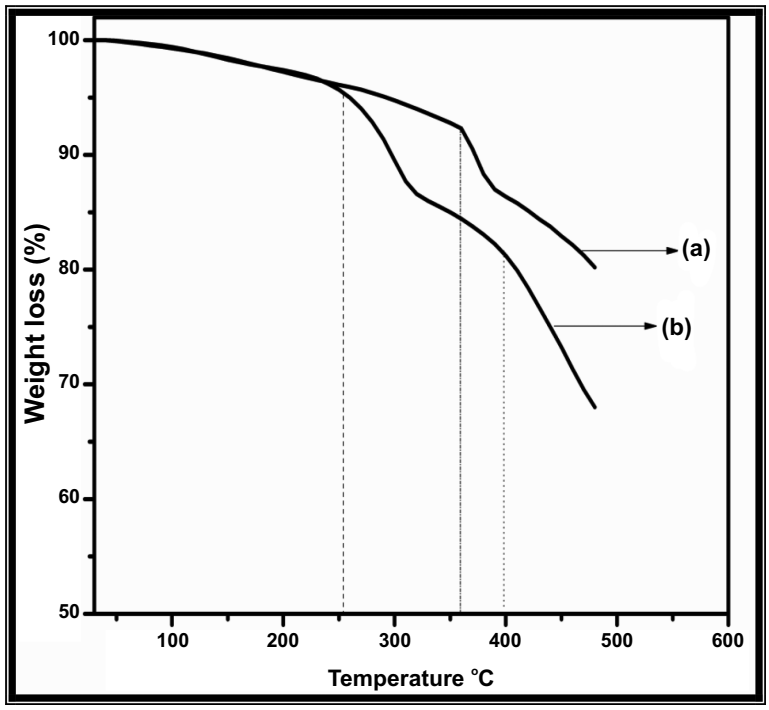


Fig. (6). TGA Thermogram of PAANa-starch/PANI (**a**) and PAANa-starch (**b**) hydrogel.

crosslinker, monomer concentration and particle size were optimized. It was found that the APS/aniline ratio could effectively be used to synthesize PANI within the matrix hydrogel.

3.8. Effect of Crosslinker Amount

Different amount of crosslinker was calculated to find out the optimum parameter of crosslinker (0.02-0.22 wt%) necessary for forming hydrogel network. The conductivity of PAANa-starch/PANI with a varying concentration of crosssslinker was evaluated to find out the optimum concentration of crosslinker. Table **1** shows the data for the development of conductivity in swelled sample of PAANa-starch/PANI hydrogel as the proportion of crossslinker was increased. The conductivity shows a sharp increase even with the incorporation of 0.08 wt% of crossslinker amount. With the increase in crossslinker amount, conductivity shows a smooth increase till it reaches to 2.31 mS cm^{-1} with a crossslinker concentration of 0.18 wt%. Above this concentration, there was a slow increase in the conductivity, which may be due to the enhancement in the degree of crosslinking which consecutively lowers the flexibility of the polymeric chains [35]. Thus, it reduced the mesh size and decreased the internal surface area of the networks [36]. Hence, the storage of a sufficient amount of aniline hydrochloride into the matrix is decreased. Further, beyond an optimum ratio of crosslinker NMBA (0.18 wt%) leads to a slow increase in conductivity value due to excessive crosslinking resulting in the formation of compactness and rigidity in polymer chains.

3.9. Effect of APS Concentration

The influence of the initiator (APS) concentration in reaction mixture on the electrical conductivity of synthesized hydrogel was examined. It has been found (Table **2**) that concentration of both aniline and APS was found to possess an important role for the successful synthesis of conducting hydrogel with maximum conductivity. It has been observed that with increasing initiator amount, the conductivity is increasing. This may be due to the fact that the lower amount of APS will not form proper conducting channel inside the hydrogel matrix, which shows low conductivity value from their optimum concentration. However, after attaining an optimum concentration (8.7 x 10^{-3} mol L^{-1}), maximum conductivity value was found and, after that it was decreasing on further increase in concentration (Table **2**). This trend of conductivity may be because with the excessive amount of APS concerning aniline leads to the over oxidation of aniline chain, resulting in the formation of insulating black polyaniline [37].

Table 1. Effect of crosslinker on the conductivity of the PAANa-starch/PANI hydrogel.

S. No.	Crosslinker Amount (wt %)	Conductivity (mS cm^{-1})
1	0.02	0.23
2	0.04	0.79
3	0.08	1.10
4	0.14	1.72
5	0.18	2.31
6	0.20	2.54
7	0.22	2.61

Table 2. Effect of APS concentration on the conductivity of the PAANa-starch/PANI hydrogel.

S. No.	APS Concentration (mol/L)	Conductivity (mS cm^{-1})
1	$4.3x\ 10^{-3}$	1.26
2	$8.7\ x\ 10^{-3}$	2.09
3	$1.31\ x\ 10^{-2}$	1.68

3.10. Effect of Aniline Concentration

The change in electrical conductivity of hydrogel samples concerning increase in aniline concentration are shown in Table **3**. It is observed that the conductivity of the PAANa-starch/PANI hydrogel increases with an increase in the concentration of aniline from (0.5-2.75 wt %), but after attaining an optimum concentration, the conductivity of the PAANa-starch/PANI hydrogel retards. The higher amount of aniline resulted in the poor distribution of aniline followed by polymerization within the matrix, or it gets over-oxidized soon giving a lower conductivity value of hydrogel. On the other hand, an increase in the concentration of aniline provides a presence of an excess number of aniline molecules that are absorbed in the hydrogel matrix which increases the electrical conductivity of hydrogel. However, further increase in the concentration of aniline after attaining its optimum concentration (2 wt%); there is a slow increase in the conductivity of the hydrogel sample. This could be due to the occurrence of more ionic concentration (presence of aniline hydrochloride ion) in the external solution, which results in the smaller amount of absorption of aniline molecule inside the hydrogel matrix and restricts the proper amount of aniline for the polymerization within the matrix [38]. Hence, it evident that an adequate ratio of aniline and APS has been taken for smooth progress of the IPN gel in good condition (considering homogeneity, incorporation, and electrical conductivity) and to avoid instant surface polymerization of the hydrogel matrix, its surface hydrogel should be lightly washed. Hence, the significance of the optimization to get the synthesized product with high electrical conductivity along with proper synthesis condition of PANI into the matrix hydrogel was realized.

Table 3. Effect of aniline concentration on the conductivity of the PAANa-starch/PANI hydrogel and their filtrate.

S. No.	Aniline Concentration (wt %)	Hydrogel Conductivity (mS cm^{-1})	Filtrate Conductivity (mS cm^{-1})
1	0.50	1.12	0.15
2	1.00	2.31	0.67
3	1.50	2.91	0.82
4	2.00	3.21	0.89
5	2.50	3.59	0.91
6	2.75	3.61	0.93

The comparison of hydrogel with their resultant filtrate (Table **3**) shows that the conductivity of hydrogel samples is more than their filtrate solution. It implies that the conductivity is occurring due to the presence of PANI chains within the hydrogel matrix.

3.11. Comparison of the Conductivity of P(AANa-co-AM)/PANI with PAANa-Starch/PANI Hydrogel

The comparison of electrical conductivity of P(AANa-co-AM)/PANI with PAANa-starch/PANI synthesized composite hydrogel at an optimum concentration of crosslinker, initiator, and aniline has been performed. The highest conductivity is obtained in PAANa-starch/PANI hydrogel (3.59 mS cm^{-1}) sample than P(AANa-co-AM)/PANI hydrogel (0.91 mS cm^{-1}). This enhancement in electrical conductivity in case of PAANa-starch/PANI hydrogel could be due to the formation of more crosslinks of synthetic polymer polyacrylamide in P(AANa-co-AM)/PANI hydrogel system which restricts the proper distribution of aniline hydrochloride solution inside the P(AANa-co-AM) hydrogel matrix. This hindrance caused poor polymerization of aniline, resulting in a decrease in the hydrogel conductivity.

3.12. Effect of Particle Size of SAP on Electrical Conductivity

The influence of particle size on the conductivity of the hydrogel sample is also studied. The hydrogel prepared with 40 particle size shows conductivity value in the range of 3.37 mS cm^{-1} while with 60 particle size, the conductivity value is 3.59 mS cm^{-1} at an optimum composition of other ingredients. It was found that the particle size does not influence the diffusion and electrical conductivity of the synthesized hydrogel sample much, as the conductivity of hydrogel is generally generated due to the presence of conducting material PANI within the matrix as hydrogel is insulating in nature.

3.13. Effect of Starch Content on the Synthesis of SAP Hydrogel Matrix

The influence of starch content on the synthesis of the matrix has been studied by varying the amount of starch in the range of 0.01 to 0.03 wt%, keeping the other compositions constant. It is observed that a lesser amount of starch does not bind with crosslinking network structure, and therefore it results in improper crosslinking hydrogel network structure which can hold the aniline hydrochloride molecule inside the matrix to form high conducting hydrogel; while, if the quantity of starch is beyond 0.2 wt%, then it will not disperse in monomer solution to form hydrogel with proper crosslinking point. This could be due to the occurrence of a high degree of polymerization with additional crosslinks points within the polymer networks.

CONCLUSION

A conducting hydrogel based PAANa-starch/PANI was successfully prepared *via* aqueous solution polymerization by optimizing the synthesis parameters. The effect of preparation conditions on the electrical property has also been investigated. It is seen from the data that the PAANa-

starch/PANI hydrogel exhibits better conductivity (3.59 mS cm^{-1}) than P(AANa-co-AM)/PANI hydrogel (0.91 mS cm^{-1}). This enhancement in electrical conductivity of PAANa-starch/PANI hydrogel could be due to the formation of more crosslinks in P(AANa-co-AM)/PANI hydrogel system which restricts the proper distribution of aniline hydrochloride solution inside the P(AANa-co-AM) hydrogel matrix. This hindrance to monomer is causing poor polymerization of aniline and thus resulting in a decrease in the hydrogel conductivity.

LIST OF ABBREVIATION

AA	=	Acrylic Acid
APS	=	Ammonium Peroxidisulphate
COO^-	=	Carboxylate ion
COOH	=	Carboxyl
CP	=	Conducting Polymers
FTIR	=	Fourier Transform Infrared
IPN	=	Interpenetrating Polymerization
N'N,MBA	=	N'N Methylene Bisacrylamide
NaOH	=	Sodium Hydroxide
NH_2	=	Amide
P(AANa-AM)/ PANI	=	Poly(acrylate-co-acrylamide)/Polyaniline
PAAM	=	Polyacrylamide
PAANa	=	Polyacrylate
PAANa-starch/ PANI	=	Polyacrylate-starch/Polyaniline
PANI	=	Polyaniline
SEM	=	Scanning Electron Microscopy
SR	=	Swelling RATIO
TGA	=	Thermogravimetric Analysis
X-RD	=	X-Ray Diffraction

CONSENT FOR PUBLICATION

Not applicable.

AVAILABILITY OF DATA AND MATERIALS

Not applicable.

FUNDING

None.

CONFLICT OF INTEREST

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be constructed as a potential conflict of interest.

ACKNOWLEDGEMENTS

The financial assistance from Delhi Technological University (DTU), Delhi to Reetu Prabhakar is acknowledged gratefully.

REFERENCES

[1] Omidian, H.; Rocca, J.G.; Park, K. Advances in superporous hydrogels. *J. Control Rel.*, **2005**, *102*, 3-12.

[2] Li, G.; Zhao, Y.; Zhang, L.; Gao, M.; Kong, Y.; Yang, Y. Preparation of graphene oxide/polyacrylamide composite hydrogel and its effect on Schwann cells attachment and proliferation. *Colloid Surf. B: Biointerf.*, **2016**, *143*, 547-556.

[3] Li, M.Y.; Guo, Y.; Wei, Y; Mac Diarmid, A.G.; Lelkes, P.I. Electrospinning polyaniline-contained gelatin nanofibers for tissue engineering applications. *Biomaterials*, **2006**, *27(13)*, 2705-2715.

[4] Siddhanta, S.K.; Gangopadhyay, R. Conducting polymer gel: Formation of a novel semi-IPN from polyaniline and crosslinked poly(2-acrylamido-2-methyl propanesulphonicacid). *Polym. J.*, **2005**, *46*, 2993-3000.

[5] Small, C.J.; Too, C.O.; Wallace, G.G. Responsive conducting polymer-hydrogel composites. *Polym. Gel Netw.*, **1997**, *5*, 251-265.

[6] Wada, H.; Nohara, S.; Iwakura, C. Electrochemical characteristics of electric double layer capacitor using sulfonated polypropylene separator impregnated with polymer hydrogel electrolyte. *Electrochim. Acta.*, **2004**, *49*, 4871-4875.

[7] Wu, J.H.; Wei, Y.L.; Lin J.M.; Lin, S.B. Study on starch-graft-acrylamide/mineral powder superabsorbent composite. *Polymer*, **2003**, *44*, 6513-6520.

[8] Ito, K.; Chuang, J.; Carmen, A.L.; Watanabe, T.; Ando, N.A.; Yu, G. Multiple point adsorption in a heteropolymer gel and the Tanaka approach to imprinting: Experiment and theory. *Prog. Polym. Sci.*, **2003**, *28*, 1489-1515.

[9] Tang, Q.; Wu, J.; Sun, H.; Fan, S.; Hu, D.; Lin, J. Superabsorbent conducting hydrogel from poly(acrylamide-aniline) with thermosensitivity and release properties. *Carbohyd. Polym.*, **2008**, *73*, 473-481.

[10] Tang, Q.; Wu, J.; Lin, J. A multifunctional hydrogel with high conductivity, pH-responsive, thermo-responsive and release properties from polyacrylate/polyaniline hybrid. *Carbohyd. Polym.*, **2008**, *73*, 315-321.

[11] Lin, J.; Tang, Q.; Wu, J. The synthesis and electrical conductivity of a polyacrylamide/Cu conducting hydrogel. *React. Funct. Polym.*, **2007**, *67*, 489-494.

[12] Lin, J.; Tang, Q.; Wu, J.; Hao, S. The synthesis and electrical conductivity of a polyacrylate/graphite hydrogel. *React. Funct. Polym.*, **2007**, *67*, 275-281.

[13] Ali, T.O. The corrosion performance of polyaniline film modified on nickel plated copper in aqueous p-toluenesulfonic acid solution. *Surf. Coat. Technol.*, **2006**, *200*(12-13), 3918-3925.

[14] Lin, J.; Tang, Q.; Wu, J.; Sun, H.; Fan, S.; Hu, D. Two-step synthesis and conductivity of polyacrylamide/Cu conducting hydrogel. *Polym. Compos.*, **2009**, *30*, 1132-1137.

[15] Tang, Q.; Sun, X.; Wu, J.; Li, Q.; Lin, J. Design and electrical conductivity of poly(acrylic acid-g-gelatin)/graphite conducting gel. *Polym. Eng. Sci.*, **2009**, *49*, 1871-1878.

[16] Chauhan, N.P.S.; Jangid, N.K.; Punjabi, P.B. Synthesis and characterization of conducting polyaniline *via* catalytic oxidative polymerization. *Int. J. Polym. Mater. Polym. Biomater.*, **2013**, *62*, 550-555.

[17] Virji, S.H.; Weiller, B.H.; Huang, J.; Blair, R.; Shepherd, H.; Faltens, T. Construction of a polyaniline nanofibre gas sensor. *J. Chem. Educ.*, **2008**, *85*, 1102-1104.

[18] Zhang, L.; Wan, M.; Wei, Y. Nanoscale polyaniline fibers prepared by ferric chloride as an oxidant. *Macromol. Rapid Commun.*, **2006**, *27*, 366-371.

[19] Sairam, M.; Nataraj, S.K.; Aminabhavi, T.M.; Roy, S.; Madhusoodana, C.D. Polyaniline membranes for separation and purification of gases, liquids, and electrolyte solutions. *Sep. Purif. Rev.*, **2006**, *35*, 249-283.

[20] Malhotra, B.D.; Ghosh, S.; Chandra, R. Polyaniline/polymeric acid composite, a novel conducting rubber. *J. Appl. Polym. Sci.*, **1990**, *40*, 1049-1052.

[21] Tiwari, A.; Sharma, Y.; Hattori, S.; Terada, D.; Sharma, A.K.; Turner, A.P.F.; Kobayashi, H. Influence of poly(n-isopropylacrylamide)-CNT-polyaniline three-dimensional electrospun microfabric scaffolds on cell growth and viability. *Biopolymers*, **2013**, *99*, 334-341.

[22] Wu, J.; Wei, Y.; Lin, J.; Lin, S. Study on starch-*graft*-acrylamide/mineral powder superabsorbent composite. *Polymer*, **2003**, *44*, 6513-6520.

[23] Lu, S.; Duan, M.; Lin, S. Synthesis of superabsorbent starch-*graft*-poly(potassium acrylate-*co*-acrylamide) and its properties. *J. Appl. Polym. Sci.*, **2003**, *88*, 1536-1542.

[24] Li, W.; Zhao, H.; Teasdale, P.R.; John, R. Preparation and characterization of poly(2-acrylamidoglycolic acid-co-acrylamide) hydrogel for selective binding of Cu^{2+} and application to diffuse gradients in thin films measurements. *Polymer*, **2002**, *43*, 4803-4809.

[25] Patil, R.C.; Patil, S.F.; Mulla, I.S.; Vijayamohanan, K. Effect of protonation media on chemically and electrochemically synthesized polyaniline. *Polym. Int.*, **2000**, *49*, 189-196.

[26] Josefowicz, M.E.; Epstein, A.J.; Tang, X. Protonic acid doping of two classes of the emeraldine form of polyaniline. *Synth. Met.*, **1992**, *46*, 337-340.

[27] Mahdavinia, G.R.; Pourjavadi, A.; Hosseinzadeh, H. Modified chitosan 4. Superabsorbent hydrogels from poly(acrylic acid-co-acrylamide) grafted chitosan with salt- and pH-responsiveness properties. *Eur. Polym. J.*, **2004**, *40*, 1399-1407.

[28] Zhao, Y.; Su, H.; Fang, L.; Tan, T. Superabsorbent hydrogels from poly(aspartic acid) with salt-, temperature- and pH-responsiveness properties. *Polymer*, **2005**, *46*, 5368-5367.

[29] Prabhakar, R.; Kumar, D. Influence of dopant ions on the properties of conducting polyacrylamide/polyaniline hydrogels. *Poly. Plast. Technol.*, **2016**, *55*, 46-53.

[30] Tzou, K.; Gregory, R.V. A method to prepare soluble polyaniline salt solution *in situ* doping of PANI base with organic dopants in polar solvents. *Synth. Met.*, **1993**, *53*, 365-377.

[31] Ibrahim, K.A. Synthesis and characterization of polyaniline and poly(aniline-co-nitroaniline) using vibrational spectroscopy. *Arab. J. Chem.*, **2017**, *10*, 2668-2674.

[32] Zang, H.; Lu, J.; Wang, X.; Li, J; Wang, F. From amorphous to crystalline: Practical way to improve electrical conductivity of waterborne conducting polyaniline. *Polymer*, **2011**, *52*, 3059-3064.

[33] Han, M.; Chu, Y.; Han, D.; Liu, Y. J. Fabrication and characterizations of oligopyrrole doped with dodecylbenzenesulfonic acid in reverse microemulsion. *Colloid Interf. Sci.*, **2006**, *296*, 110-117.

[34] Sharma, K.; Kaith, B.S.; Kumar, V.; Kalia, S.; Som. S.; Swart, H.C. Gum ghatti based novel electrically conductive biomaterials: A study of conductivity and surface morphology. *eXPRESS Polym. Lett.*, **2014**, *8*, 267-281.

[35] Tiwari, A.; Singh, V. Microwave-induced synthesis of electrical conducting gum acacia-graft-polyaniline. *Carbohyd. Polym.*, **2008**, *74*, 427-434.

[36] Wu, J.; Lin, J.; Zhou, M.; Wei, C. Synthesis and properties of starch-graft-polyacrylamide/clay superabsorbent composite. *Macromol. Rapid Commun.*, **2000**, *21*, 1032-1034.

[37] Siddhanta, S.K.; Gangopadhyay, R. Conducting polymer gel: Formation of a novel semi-IPN from polyaniline and crosslinked poly(2-acrylamido-2-methyl propanesulphonicacid). *Polymer*, **2005**, *46*, 2993-3000.

[38] Bajpai, A.K.; Bajpai, J.; Soni, S.N. Preparation and characterization of electrically conductive composites of poly(vinyl alcohol)-g-poly(acrylic acid) hydrogels impregnated with polyaniline (PANI). *eXPRESS Polym. Lett.*, **2008**, *2*, 26-39.

International Journal of Research in Engineering and Innovation (IJREI)
journal home page: http://www.ijrei.com
ISSN (Online): 2456-6934

Thermodynamic analysis of combined cycle power plant with trans-critical cycle integrated with solar system for power generation for space heating and cooling

R.S. Mishra, Aakash Behl

Department of Mechanical & Production Engineering, Delhi Technological University, Delhi India

Abstract

To build up a programming code and solving it on computational numerical technique by developing energy balance equation for combine cycle power plant integrated with Trans-critical cycle. A computer programming code has been used to simulate a performance under various different conditions under different temperature and pressure circumstances and thermal performance on different cases of Brayton cycle have been analyzed using standard fluids. In this paper Brayton cycle, Rankine cycle and trans-critical cycle are combined which makes it tri-generate cycle and variations of temperature and pressure with thermal performance parameters such as thermal efficiency, exergy destructions in the various components using trans-critical cycle have been analyzed. It is observed that by increasing condenser temperature of Rankine cycle, the combined cycle efficiency , net-work are increasing and by increasing the mass flow rate of Brayton cycle, combined cycle efficiency is decreasing while combined net-work is increasing. Similarly by increasing the exit pressure of pump in the trans-critical cycle, both the combined cycle efficiency & combined net work are decreasing also is decreasing and also by increasing the inlet temperature of turbine of trans-critical cycle, the combined cycle efficiency and combined net-work is increasing. By increasing the inlet pressure of pump of trans-critical cycle, combined cycle efficiency and combined net work (i.e. both) are increasing. . © 2018 ijrei.com. All rights reserved

Keywords: Thermodynamics, Combined cycle power plant, integrated solar system

1. Introduction

The Combined cycle power plants also calls as CCPT are the combination of a gas turbine or Brayton cycle with Rankine cycle. They are designate by the high efficiency in the technology of production of electricity from fossil fuels. Simultaneously, these units are in between to the fast developing currently achieving efficiencies around 60%. Combined cycle power plants defines a number of benefits, listed as speedy construction time, mild investment costs, highly reliable and flexible. The CO_2 emission at the efficiency 0.60 is equal 330 kg--CO_2/MWh, i.e. 2 times lower than the emission from modern coal power plants (800 kgCO_2/MWh). Presently, this technology is not famous because of the high price of gas. The electric efficiency of the Brayton cycle mainly relies on the pressure ratio in the air compressor and the highest temperature in the circuit that is a Combustor Outlet Temperature (COT) of the flue gases from the combustion chamber. However, often because the most vital temperature within the rotary turbine shall be thought of the common exhaust gas temperature at the recess to the turbine, Turbine Recess Temperature (TRT).

Combined cycle power plant with the trans-critical cycle is explained here. With its combination with CO_2 trans-critical cycle it's become tri generative cycle. It has been applied in refrigeration (commercial) and heat pump water heating systems. Its cycle COP (coefficient of performance) mainly depends on high pressure optimization and control. Theoretically, the optimal high pressure increases with the increasing gas cooler outlet temp or ambient temperature. However, due to reliability and cost concerns, there is an upper limit of the high pressure in compressor design and system operation. In this work, i proposed a constrained optimization method for getting constrained optimal high pressure equation of CO_2 trans-critical cycle.

Corresponding author: R. S. Mishra
Email Id: hod.mechanical.rsm@dtu.ac.in

1.1 Combined cycle power plant

Combined cycle power plant works on two heat engines that is Brayton cycle and Rankine cycle. In order to explain the complete combine cycle power plant one need to explain the Rankine cycle and Brayton cycle separately and trans-critical cycle is also added.

1.2 The Rankine cycle

The Rankine cycle thoroughly describes the way through which team-operated heat engines largely found in thermal power generation plants for generating power. The sources used in these power plants are sometimes fission or the combustion of the fuels accessible. The power of the Rankine cycle is prohibited by the high heat of vaporization of the operating fluid. Also, unless the pressure and temperature reach to their super critical levels within the vessel, the temperature vary the cycle will operate over is sort of small: turbine entry temperatures air usually around 565°C and steam condenser temperatures air around 30°C. This low rotary turbine entry temperature (compared to a rotary turbine) is why the Rankine (steam) cycle is usually used as a bottoming cycle to recover otherwise rejected heat in combined-cycle gas turbine power plant as additionally called gas turbine power stations.

In Rankine cycle working fluid usually circulated in a close loop and keeps recycled at every initial stage. The condensed droplets of water vapour coming out form the Rankine cycle power station through the cooling tower system not actually forms closed loop. In further step addition of steam at high temperature is required for which departure of waste product form the cycle is necessary and all this happens in a closed cycle. Generally, Q_{out} is used to represent the heat coming from the exit point, it is also called as exhaust heat .Latent heat of vaporization of the working fluid is being absorbed by the cooling towers which works as a heat exchangers and simultaneously evaporating cooling water to the atmosphere. There are many substances present which can be used as working fluid but mainly water is considered as a working fluid as it is nontoxic, easily available.

Basically there are totally four process in the working of Rankine cycle which shown below in the T-s diagram of the Rankine cycle in the below parts shown in fig.2.

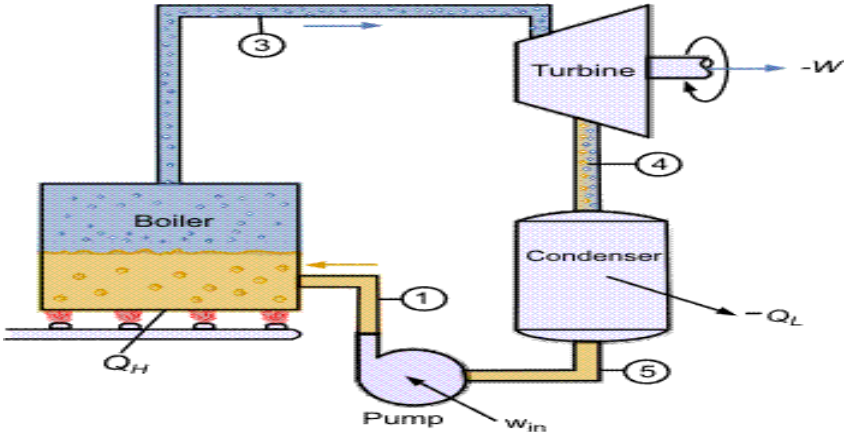


Figure 1: Basic design of Rankine cycle [1].

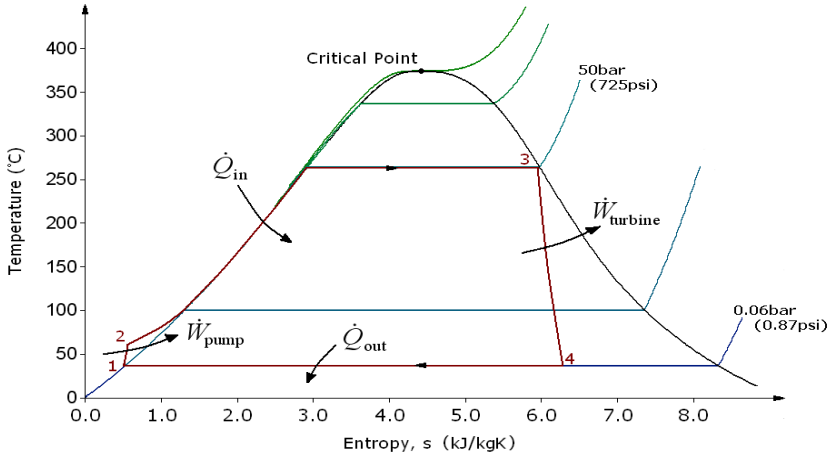


Figure 2: T-s diagram for Rankine cycle.

Process 1-2

In this process a pump is used which is shown in the Fig.2. Here pump is used to produce high pressure to the working fluid so that it can reach up to the boilers.

Process 2-3

As the working fluid reached boiler through pump, it starts getting heated through some external source so as to make working fluid as a dry saturated vapour for calculating required input energy.

Process 3-4

In this process turbine is used where expansion of dry saturated vapour takes place for the generation of power. Occurrence of condensation along with depletion of temperature and pressure is seen. Numerically using the steam tables one can calculate the outputs.

Process 4-1

Here on this process condenser is used where saturated liquid enters the condenser to condense at constant pressure.
In a perfect temperature cycle, the pump and rotary engine will be like physical property, that is no entropy will be generated by either pump or rotary engine and thus maximize output work. On T-s diagram (Fig.2), the vertical lines are used to define the process 1-2 as well as process 3-4 and additionally closely match that of the Carnot cycle. In Fig.2, it is shown that prevention of vapour ending with in the superheat region in temp unit cycle.

1.3 Brayton Cycle

The Brayton cycle is the ideal cycle for gas-turbine engines. It is used for gas turbines in which both the compression and expansion processes are implemented. There are two different types of the Brayton cycle; the open gas turbine cycles and the closed gas turbine cycle respectively. The difference between these two cycles is that during the open gas turbine cycle, a combustion process takes place, and exhaust gases are thrown out, in other words the exhaust gases cannot be recirculates, while in the other cycle (the closed gas turbine cycle), the combustion process is replaced by a heat-addition process, the exhaust gases are also utilized so as to increase the temperature of the air which enters the compressor. In Fig.3, we can see an open gas turbine. Initially, at ambient condition fresh air is taken up by the compressor, and here as a result of compressions the temperature and pressure is increased.. Here, the combustion process occurs at constant temperature. In the third process the high temperature gas enters in to turbine as you can see in Fig.3. In this operation, at atmospheric pressure gases (hot) will expand in the turbine so that power generation can take place. At last exhaust gases coming out from the turbine will move back to the heat exchanger to atmosphere.

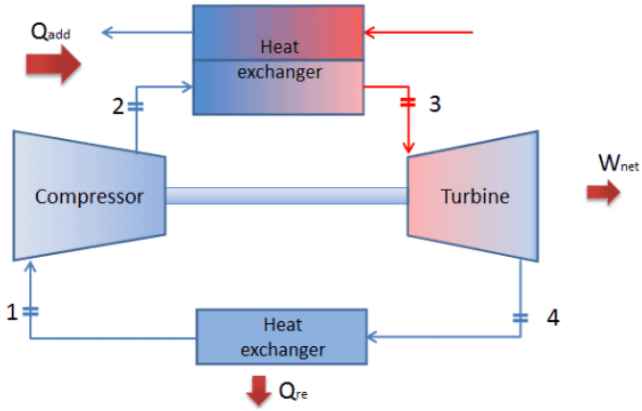


Figure 3: The Brayton cycle [3]

In the shut gas turbine cycle, in spite of the fact that the pressure and development process have in like manner, burning chamber is supplanted by a warmth exchanger in which builds the compacted air temperature. As given in Fig.3, perfect Brayton cycle is really a shut gas turbine cycle, and the means of the Brayton cycle resemble following processes.

Process 1–2

It is an isentropic compression in the compressor, in this process a pump is used which is shown in the Fig.3. Here pump is used to produce high pressure to the working fluid so that it can reach up to the boilers.

Process 2–3

It is known as constant pressure heat addition. As the working fluid reached boiler through pump, it starts getting heated through some external source so as to make working fluid as a dry saturated vapour. Graphically and numerically we can calculate the required input energy.

Process 3–4

It is an isentropic expansion in the turbine. In this process, the turbine is used where expansion of dry saturated vapour takes place for the generation of power and occurrence of condensation along with depletion of temperature and pressure is observed. Numerically using the steam tables one can calculate the outputs.

Process 4–1

It is constant pressure heat rejection, here on this process condenser is used where saturated liquid enters the condenser to condense at constant pressure.

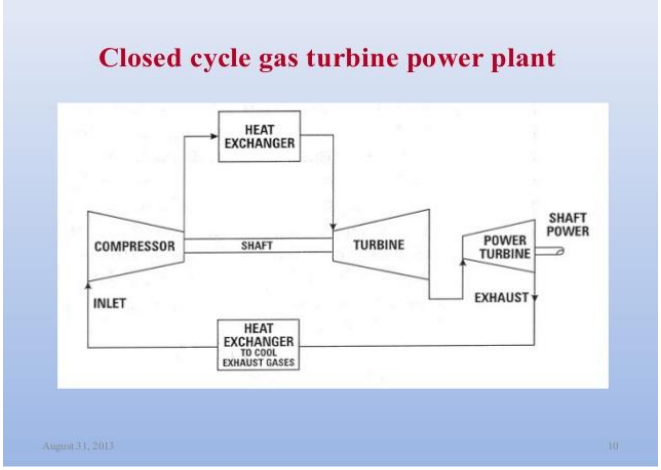


Figure 4: A closed-cycle gas turbine engine [4]

Gaseous petrol as fuel is utilized to run an open Brayton cycle, flammable gas is copied with compacted air to deliver high weight hot vent gases having temperature over 1000ºC. These gases extends in the turbine driving the generator and the fumes gases are utilized as a fuel in squander warm recuperation kettle or warmth recuperation steam generator. The superheated steam at that point drives the turbine therefore creating extra power alongside the power delivered from Brayton cycle. Once in a while the turbine from both the cycles is in single shaft and they together drive the generator to deliver power, for the most part the two turbines have their own particular generator. The gas turbine is a standout amongst the most productive one for the change of gas powers to mechanical power or power. The utilization of distillate fluid powers, generally diesel, is likewise normal as exchange energizes. All the more as of late, as straightforward cycle efficiencies have enhanced and as petroleum gas costs have fallen, gas turbines have been all the more generally embraced for base load control age, particularly in consolidated cycle mode, where squander warm is recouped in squander warm boilers, and the steam used to deliver extra power. They are installed around the world. Consolidated cycle control plant as in name recommends, it joins existing gas and steam innovations into one unit, yielding noteworthy changes in warm proficiency over traditional steam plant. Anyway the warmth recouped in this procedure is adequate to drive a steam turbine with an electrical yield of roughly 50 for every penny of the gas turbine generator. The gas turbine and steam turbine are coupled to a solitary generator. For start-up, or 'open cycle' task of the gas turbine alone, the steam turbine can be detached utilizing a pressure driven grasp. As far as general speculation a solitary shaft framework is regularly around 5 for every penny bring down in cost, with its working straightforwardness ordinarily prompting higher unwavering.

1.4 Carbon dioxide Trans-critical Power Cycle

The current study conjointly focuses on the trans-critical cycle due to its high potential usage within the trade and thanks to the restricted studies found within the literature. The trans-critical cycle, whose heat rejection takes place at a subcritical pressure, should not be confused with the entirely critical cycle.

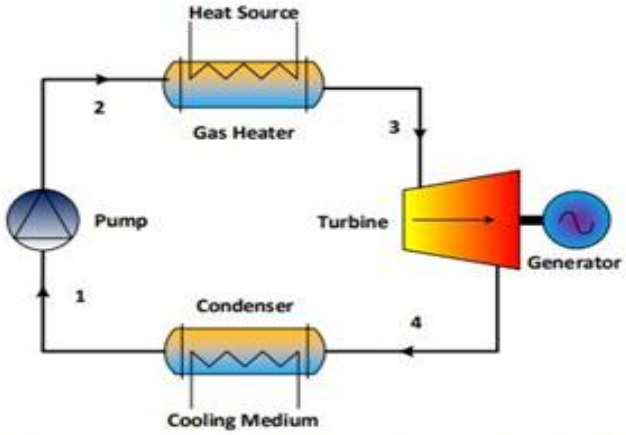


Figure 5: The trans-critical cycle [5]

Actually, coal pink-slipped trans-critical power plants at high temperatures (above five hundred °C) represent a mature technology and among the most effective playing heat engines with a thermal potency as high as forty ninth .As way because it is understood the dioxide are going to be thought-about as trans-critical cycle wherever the temperature is on top of vital temperature i.e., 31 °C. For trans-critical carbon dioxide because it is portrayed in Fig.6 and therefore the T-S diagram in Fig.6 same as Brayton cycle the trans-critical dioxide cycle can expertise processes: compression (1-2), isobaric heat provide (2-3), growth (3-4), and isobaric heat rejection (4-5). The sole distinction between these 2 cycles is whether or not a part of the cycle is found within the subcritical region or not. Therefore, each cycle' generally associated with critical cycles within the literature.

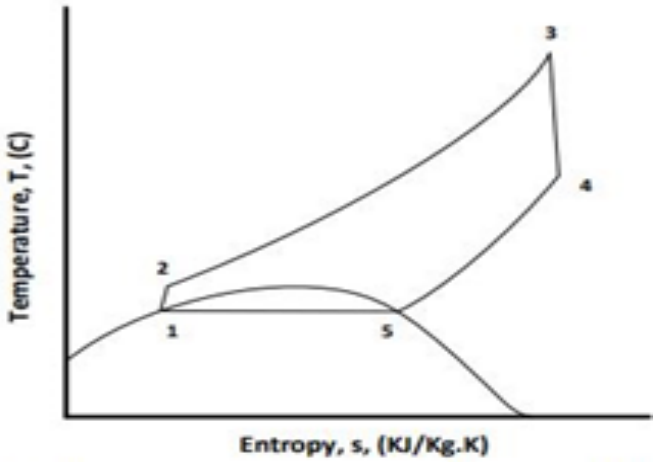


Figure 6: T-S Diagram of Trans-critical Cycle

The energy analysis relies on the primary law of physics. The thermal potency and therefore the specific internet output ar its results. With the assumptions antecedent explicit, their values rely solely on one freelance parameter: the air mass, that ar P2=P3. Above all, these results don't depend upon the operating fluid mass rate. The equations for the various parts. For the pump: $\eta_p = (h_2,s-h_1)/ (h_2-h_1)$ wherever the η_p is that the potency of the pump, h_1 is that the specific physical property of the pump recess fluid, h_2 is that the physical property of pump outlet fluid and $h_{2,s}$ is that the physical property physical property of outlet fluid. $w_p = h_2 - h_1$ wherever the w_p is that the work of the pump, $h_1$is the particular physical property of the recess fluid and h2 is that the physical property of outlet fluid. For the turbine: $\eta_t = (h_3-h_4)/ (h_3-h_{4,s})$ wherever the η_t is that the potency of the rotary engine, h3is the particular physical property of carbon dioxide at the rotary engine recess, h_4 is that the physical property of outlet fluid and h4,is is that the physical property physical property of outlet fluid. $w_t = h_3 - h_4$ wherever the w_t is that the work of the pump, $h_3$is the particular physical property of the rotary engine recess fluid and h_4 is that the specific physical property of outlet fluid. For the vapor generator: $q_{in} = h_3 - h_2$ wherever the q_{in} is that the heat transferred to the fluid in vapor generator, h_2 is that the specific physical property of the warmer recess fluid and h_3 is that the physical property of outlet fluid. For the condenser: $q_{out} = h_4 - h_1$ wherever the $q_{_out}$ is that the heat rejected from the operating fluid in condenser, h4is the physical property of the fluid coming into the condenser and h1 is that the physical property exiting of fluid deed the condenser. The thermal potency of the cycle:

$$\eta_{th} = [(W_t-W_p)/ q_{_in}] = [(h_3-h_4)-(h_2-h_1)/ (h_3-h_2)]$$

wherever the $\boldsymbol{\eta_{th}}$ is that the thermal potency of the cycle, w_t is that the rotary engine work, W_p is that the pump work and q_{in} is heat transferred to the operating fluid in warmer. In compare with associate degree organic Jeannette Rankin cycle (ORC), the carbon dioxide trans-critical power cycle encompasses a higher capability in taking blessings of the energy during an inferior waste heat with gradient temperature, like exhaust gases. The temperature glide (Temperature modification throughout take-up of warmth energy) for carbon dioxide on top of the crisis permits for an improved.

1.5 Turbine Cycle

An ignited mixture of a natural gas and air which is purified and a later on compressed makes it to expand. Spinning of a turbine blades took place due to a pressure created form expansion process, for creating electricity ay are attached with a generator as well as shaft. In further second step a generation of steam is done by utilizing a heat of gas turbine from exhaust and make it to through a heat recovery steam generator (HRSG).

1.6 Heat Recovery Steam Generator

In HRSG that stands for Heat Recovery Steam Generator is known for water flow in a tube and a flow is extremely purified and around am passing of a hot gases occurs and therefore ay produce steam .In a steam turbine rotation of steam occurs and concatenate with a generator for a production of Electricity.A steam condensing and water system is a same as in a steam power plant.

1.7 Typical Size and Configuration of CCGT Plants

Rankine and Brayton cycle combines to form a combined-cycle. It includes different configuration for shafts such as multi shaft configuration and single shaft configuration. Single-shaft systems constitute of one - one unit of each gas turbine, steam turbine, generator and Heat Recovery Steam Generator (HRSG), with a gas turbine and steam turbine concatenate to a single shaft of a single generator. Areare one or more generator in multi turbines and HRSGs to a separate single steam turbine-generator steam is supplied through it.
A chief demerit of multiple stage combined cycle power plant is that a number of steam turbines, condensers and condensate systems-and perhaps a cooling towers and circulating water systems increases to equalise a quantity of gas turbines.

1.8 Efficiency of CCGT Plant

It is essential to use a parts of a exhaust energy through gas to gas healer after a combination of Rankine cycle and a bray ton cycle, achievement of low temperature at output and high temperature at a input can be done. A efficiency of a cycles will sum up with it, as same fuel source powered am. A power of CCPT is filled by Supplementary Firing and Blade Cooling. At HRSG supplementary firing was organised and a flow of compressed gas flow is being carried out in a turbine area of a rotary engine and it was mainly opted for a cooling down of a turbine blades.

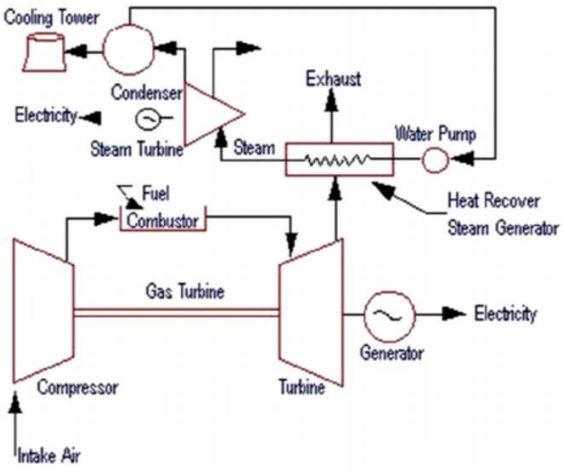


Figure 7: Combined Cycle Gas Turbine Plant HRSG [6]

1.9 Combined cycle integrated with solar system

PTS called as a parabolic trough solar energy plant. Ay are a most vital establishment of a CSP that is concentrating solar energy technique. SEGS that is PT star electricity producing technique system shows that a tendency of this present technique found to be more trust worthy as well as renewable resource. This method usually adopted on a commercial scale on large scale with some around output of 345MW. For satisfying a increasing demand for a energy resource, CSP are found to be a most promising technology. It has been seen that only CSP can fulfil a demand of energy for developing countries. Electricity produce by a CSP with in a MENA (Mediterranean & geographical region) wants to improvise a native energy production system. Trans Mediterranean renewable energy cooperation introduced a energy cooperation. ISCC is one of a most important system. It stands for integrated solar combine cycle; PTT that is parabolic trough technology is said to be coupled whenever it is supported by a star field. Are is no need to install a storage source for a backup of fuel for boilers because of a flexibility in operation and reduction in value of a system. Here a mathematical codes are built for a integrated solar combine cycle that is ISCC powerhouse operational below weaar conditions. Habib (1994)[1] presented an analysis of a cogeneration system, which quantified the irreversibility of the different components of each plant. In addition to that, the influence of the heat-to-power ratio and the process pressure on the thermal efficiency and utilization factor is presented. His results show that the total irreversibility of the cogeneration plant is 38 percent lower compared to the conventional plant. This reduction in the irreversibility is accompanied by an increase in the thermal efficiency and utilization factor by 25 and 24 percent, respectively. The results show that the exergy destruction in the boiler is the highest. Bejan, et al (1996)[2] actually provided a comprehensive and rigorous introduction to thermal system design and optimization from a contemporary perspective. The book includes current developments in engineering thermodynamics, heat transfer, and engineering economics relevant to design. The use of exergy analysis and entropy generation minimization is featured. A detailed description of engineering economics and thermo economics are also presented. Moreover, a case study is considered throughout the book for continuity of the presentation. The case study involves the design of a gas turbine cogeneration system. Rolf Kehlhofer (1997) [3] provided the study of thermodynamic principles of combined cycle power plants and co-generation system. His work includes the different layout system of combined cycle power plants. It describes the effect of operating variables and part load behavior of combined cycle power plants. Karthikeyan et al (1998) [4] derived energy balances for a one pressure level heat recovery steam generator. Effects of pinch and approach points on steam generation and also on temperature profiles across heat recovery steam generator was investigated. The effects of operating conditions on steam production and also on exit gas temperature from the heat recovery steam generator are discussed. It is concluded that low pinch point results in improved heat recovery steam generator performance due to reduced irreversibility's. Additionally, the supplementary firing enhances the steam production. Tawney, et al. (2000)[5] focused on several ranges of process steam flows and conditions in order to provide a basis for comparison of the most common cycle configurations in combined cycle applications. Plant design, cycle performance, and economics of each configuration are evaluated based on requirements of flexibility and process steam flows. Rather than self-establishing the energy balances, Gate Cycle TM Heat Balance software developed by GE Enter Software, Inc. is used to build thermal models. Additionally, a financial software tool developed within Bechtel is used to construct an economic model for each cycle configuration. It is concluded that, the selection of a cogeneration facility type and the economic parameters are very much site specific and are based on numerous variables such as site ambient conditions, the level of desired power output and steam demand, capacity factor, flexibility, power purchase agreement and steam purchase agreement requirements, and owner's economic parameters for return on equity. Boyce (2002) [6] covered all major aspects of power plant design, operation, and maintenance. It covers cycle optimization and reliability, technical details on sizing, plant layout, fuel selection, types of drives, and performance characteristics of all major components in a cogeneration or combined cycle power plant. Comparison of various energy systems, latest cycles and power augmentation techniques, reviews and benefits of latest codes, detailed analysis of available equipment, techniques for improving plant reliability and maintainability, testing and plant evaluation techniques, and advantages and disadvantages of fuels. ET. Bonataki and K.C. Gianna Koglou (2002) [7], gave a modern optimization methods based on evolutionary algorithms and game theory. They will be supported by computational methods for their thermal analysis and simple model for computing there capital cost. In this paper a detailed thermal model and a quite simple economic model will be incorporated in an evolutionary search algorithm and an automated tool combine the design of new cycle gas turbine power plant will be demonstrated. The search algorithm is used a Genetic Algorithm (GA) for single and multi-objective optimization problems. The latter is handled through game theory inspired enhancement to the G.A. based search yielding the so called optimal Pareto front, with two objectives (via electrical efficiency and investment cost), the Pareto front members can be envisaged as compromise between high cost/high efficiency and low cost/low efficiency design. Neil Petchers (2003) [8] provided a comprehensive details of thermal system design from a contemporary perspective. First part of the book provide a theoretical basis for understanding the inter relation of heat and power resources. It 18 provides and introduction to basic heat and power thermodynamic and includes sections on heat and power generation technologies and equipment. Part seconds describe the infrastructure in which the theory and technologies describe in part first must be applied. Having learned on the

theory and available technology, application cannot be effective device, analysis' for cost effectiveness and implemented without knowledge of environmental factors and utility rate structures. Yongjun Zhao, et al (2003) [9], investigated the design and cost of HRSG system and to demonstrate impact on the overall cost of electricity (COE) of a combined cycle power plant. There are numerous design parameter that can affect the size and complexity of the HRSG, and it is the plan for the project to identity all the important parameter and to evaluate each the exhaust gas pressure drop across the HRSG is chosen for evaluation. This parameter affects the performance of both the gas turbine and steam turbine and size of HRSG. Single pressure, two pressure, and there pressure HRSG are investigation with the trade-offs between design point size, performance and cost evaluated for each system. A genetic algorithm is used in the design optimization process to minimize the investment cost of the HRSG second system level metrics' are employed to evaluate a design. They are gas turbine net power, steam turbine net power, fuel consumption of the power plant, net cycle efficiency of the power plant, HRSG investment cost, total investment cost of the power plant and the operating cost measured by the cost of electricity (COE), The impacts of HRSG exhaust gas pressure drop and system complexity on these system level metrics are investigated. Zaleta-Aguilar Alejandro (2003)[10],represented the proposed exergo-economic fuel-impact models for steam turbines in power plants. They are applied to calculate the impact on the steam cycle when malfunctions are occurring during the operation of steam turbine sections. Concepts such as the exergetic consumption and the dissipation temperature are used to understand the proposed fuel-impact analysis. In order to validate these fuel-impact methods, well-known procedures, to simulate on and off-design conditions of a steam power cycle are used as references. Chih Wu, (2004) [11] provided an intelligence computer software called "cycle pad". It is powerful, mature, user friendly package developed to simulate thermodynamic devices and cycles. It makes feasible for engineers to run meaningful sensitivity analysis to consider combinations of design modifications to make engineering cost benefit analysis and to include refinements such as accounting for pressure changes and heat transfers occurring between major cycles components. Yadav, et al (2004)[12] mainly focused on to development of gas turbine related power plants such as combined 20 Mitre, et al (2005), In his paper, author evaluates the effect of operational conditions on pollutants (co, co2, so2, No) emissions levels, waste heat and waste water of a combined cycle natural gas and steam power plants. The HYSYS process simulation was used for modeling and simulation. This study clearly shown that the absolute quantity of pollutants emitted is high. Also it was possible to verify that the unit operation in the condition of minimal emissions regarding the maximum possible, and thus a reduction or elimination of such pollutants is not possible. It can be observed from this study that the ideal condition for exergy productivity is to operate with a fuel air ratio as the stoichiometric one. The first constraints to this ideal is the mechanical conditions of the turbine, which can be operate at the corresponding combustion gas exit temperature so a stoichiometric ratio in the range of (2.7-2.9) is used, and these conditions make the process viable (turbine viability) and minimize pollutants production (CO and NOx). These operational conditions are the optimal considering environmental concerns. The CO2 being a product, is maximized in the process, so there is no need to search for methodologies to minimize their production, but there is for technologies for their capture and uses parallel to the process. Xiaojun Shi and DefuChe (2007)[13]proposed an improved Liquefied Natural Gas (LNG), fuelled combined cycle power plant with a waste heat recovery and utilization system. The proposed combined cycle, which provide power output and thermal energy, consist of gas/steam combined cycle, the subsystem utilization the latent heat of spent steam from the steam turbine to vaporize LNG, the sub system that recovers both the sensible heat and latent heat of water vapour in the exhaust gas from heat recovery steam generator (HRSG), by installing a condensing heat exchanger, and the HRSG waste heat utilization sub system. The conventional combined cycle and proposed combined cycle are modeled, considering mass, energy and exergy balances for every component and both energy and exergy analysis are conducted. Parametric analysis are performed for the proposed combined cycle to evaluate the effect of several factors, such as the gas turbine temperature (TIT), the condenser pressure, the pinch point temperature different of the condensing heat exchanger and fuel gas heating temperature on the performance of the proposed combined cycle through simulation calculation. The results show that the net electrical efficiency and the exergy efficiency of proposed combined cycle can be increased by 1.6% and 2.84% than those of the conventional combined cycle 21 respectively. The heat recovery per kg of flue gas is equal to 86.27 kJ/sec. one MW of electric power for operating sea water pumps can be saved. The net electric efficiency and heat recovery ratio increase as the condenser pressure decreases. The higher heat recovery from HRSG exit flue gas is achieve at higher gas TIT and at lower pinch point temperature of condensing heat exchanger. Srinivas, et al (2008) [14].studied the optimum configuration for single pressure (SP), dual pressure (DP) and triple pressure (TP) heat steam generator (HRSG) to improve heat recovery and exergy efficiency of combined cycle. Deaerator was added to enhance efficiency and remove dissolve gases in feed water. A new method was introduced to evaluate low pressure (LP) and intermediate pressure (IP) in HRSG from local flue gas temperature to get minimum possible temperature difference in heaters instead of a usual fixation of pressures. Optimum location for deaeraor was found at 1, 3, and 5 bar respectively for SP, DP and TP in heat recovery at a high pressure (HP) of 200 bar. It is concluded that optimum pressure ratio for compressor with SP, DP and TP effects in heat recovery are 8, 10 and 12 respectively at 12000 C of gas turbine inlet temperature optimum deaerator pressure is obtained at 1.3, and 5 bar for SP, DP and TP levels respectively at steam turbine inlet pressure of 200 bar. Similarly at 200 bar of HP pressure for DP and TP, steam

reheated demands 100 bar to maximize exergy efficiency for combustion chamber. Parametric analysis exhibits that gain in efficiency from single pressure heat recovery to DP and TP recovery increasing with diminishing rate. J.Li, G.Pei, Y.Z.Li, J.Ji[15] find out that with the reducing of the Organic Rankine Cycle (ORC), the engine shaft power isn't solely determined by the enthalpy drop by the expansion method however additionally the external heat loss from the expander. Theoretical and experimental support in evaluating tiny scale expander heat loss is rare. This paper presents a quantitative study on the convection, radiation and conductivity heat transfer from a kW-scale expander. A mathematical model is constructed and valid. The results show that the external radiative or convective heat loss constant was regarding 3.2 or 7.0 W/m2K once the ORC operated around 100ºC. Radiative and convective heat loss coefficients enhanced because the expander operation temperature enhanced. Conductive heat loss because of the affiliation between the expander and also the support accounted for an oversized proportion of the warmth loss. The fitting relationships between heat loss and mean temperature distinction were established. It's recommended that low conduction material be embodied within the support of expander. Mattress insulation for compact expander may be eliminated once the operation temperature is around 100ºC. V. Lemort et al. [16] had presented the results of an experimental study administered on a model of an open-drive oil-free scroll expander integrated into an ORC operating with refrigerant HCFC-123. By exploiting the expander performance measurements, the eight parameters of a scroll expander semi-empirical model are then known. The model is ready to calculate variables of 1st importance like the mass rate, the delivered shaft power and also the discharge temperature, and secondary variables like the availability heating-up, the exhaust cooling-down, the close losses, the interior escape and also the mechanical losses. The utmost deviation between the predictions by the model and also the measurements is two hundredth for the mass rate, 5-hitter for the shaft power and 3K for the discharge temperature. The valid model of the expander is finally wont to quantify the various losses and to point how the planning of the expander can be altered to realize higher performances. Kosmadakis et al. [17] suggested the constant theoretical study of a 2 stage solar Rankine cycle for Ro distillation. The current work issues the constant study of an autonomous 2 stage solar ORC for Ro desalinization. The most aim is to estimate the efficiency similarly on calculate annual energy out there for desalinization. Aleksandra & Borsukiewicz[18]suggested pumping within Rankine cycle and created calculation based mostly result for the pumping work on ORC. Analysis has been administrated for eighteen completely different organic fluids that may be used an operating fluids within the subcritical ORC system. A trial was created to search out correlations between numerous thermo-physical properties of operating fluids, specific work and power of cycle. Kim, Y.M., C.G. Kim, and D. Favrat[19]Every year, the sun irradiates the landmasses on earth with the equivalent of 19,000 billion tons of oil equivalent (toe). Only a fraction (9 billion toe) would satisfy the world's current energy requirements. Put differently, in 20 minutes, the amount of solar energy falling on the earth could power the planet for one year. Chen, H., [20] Solar ponds are large-scale solar thermal energy collectors, which are pools filled with saltwater with a density gradient from the bottom to the top. A solar pond combines heat collection and storage. With a 20°C ambient temperature, the thermal energy obtained from the solar ponds is in the form of low-grade heat at 70 to 80°C. There are low-, medium- and high- temperature solar thermal collectors, depending on their collecting temperature.

2. System Description

2.1 Solar thermal power plants

A sun unendingly provides a vast quantity of energy. As a result of a character of this energy, which is opened up, it has to be collected and targeted to be useable. Are square measure several applications and techniques wherever solar power is utilized. In star thermal power plants, solar power is absorbed as heat that is a re-modelled into electricity. Re-modelling thermal solar power to electricity may be conducted by completely different approaches. A foremost common techniques square measure concentrating alternative energy (CSP) plants and also a star chimney. CSP techniques are: star tower, parabolic dish and parabolic trough. With a star chimney, a radiation is born-again to mechanical energy by heating a air in an air solar dish (greenhouse). A heated air is allowed to flow through a chimney settled at a center of a solar dish. A buoyancy force of a air causes flow through a chimney. A flowing air drives a rotary engine that is fastened at a doorway of a chimney to come up with electricity. A star chimney consists of a solar dish or greenhouse, high created chimney and rotary engine. A storage system will be used mistreatment this method to stay a plant engaging at night-time. A straightforward idea of its storage system is to fabricate water storage to a lower place a absorbent material plate of a star collectors. This technology blessings are; it makes use of beam and diffuse radiation therefore it's able to work throughout cloudy periods, it will work twenty four hours if a storage system is utilized, a desired materials to construct it straightforward and offered in most regions of a globe, and Are’s no want for cooling water systems, therefore it's appropriate for arid locations.

.

2.2 Concentrating Solar Power (CSP) plants

CSP plants provide high temperature energy that’s employed to work standard potency cycles like a Stirling engine. Though solar power concentrated plant square measure utilized largely for production of electricity, ay could be employed in several industrial exercise. Fig.8 depicts various practice for CSP systems. One in every foremost necessary conditions for selecting a foremost appropriate method for planned exercise is that an operative temperature.

Figure 8: CSP applications India [9]

2.3 Solar Tower

This method gives a boosted magnitude relation of radiation concentration that permits star towers to realize that for heating of an air high temp is required. Depicted in a Fig. 9. The system of tower consisting a reflectors of heliostat settled in spherical array round a star adopter. Position of a sun being tracked by a reflector to make sure leading a daylight to a adopter. Medium of transfer of heat is employed within aadopter for soak up a focused alternative energy. Rejected heat is provided to work up aarmal power station. Warmth fluid transfer within a central adopter may be air &wateer, melted salt or oils. Analysis releases this method may b accustomed workup of turbine wherever air is pressurized initial so het up within a receiver to a thousand. star tower among a evidenced CSP technologies within a field.. Air capability is ten MWe every

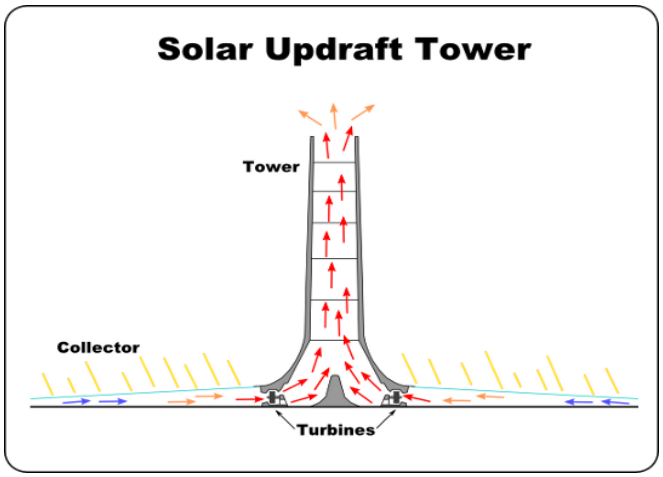


Figure 9: Solar draft (solar tower)[24]

2.4 Parabolic dish-engine

Ground level construction such method is about usage of dish having parabolic shape to focus radiations on a generator of engine within center of attention of a reflector. Under consideration of potency, umbrella dish is that a best technology among star techniques, its highest potency may be a maximum amount of twenty ninth. As you can see in fig.10. A standard umbrella dish having radius variation from five to fifteen m with five to twenty five powers. It is appropriate for redistributed potency provide and unpopulated area by taking this technique.

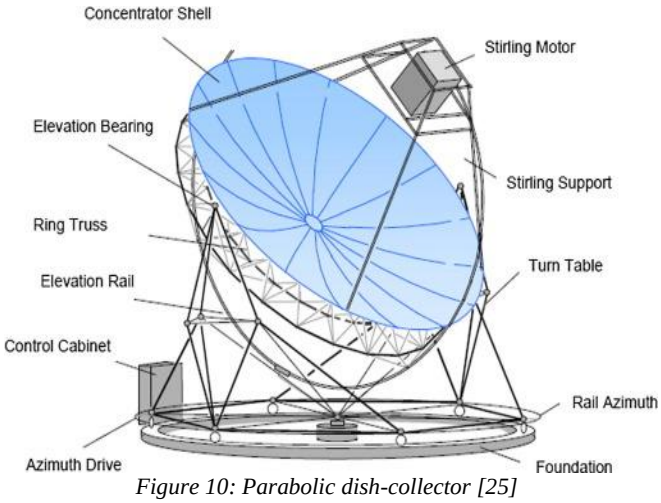


Figure 10: Parabolic dish-collector [25]

2.5 Liner Fresnel system

Setup actually have associate array of liner reflectors. A absorbent tube that is homeward on a focused radiation being

received by the focal line of the reflector and converts a sun potency to heat. Fig.11 depicts Augustin Jean Fresnel system parts. H T fluid is employed to soak up this energy to be employed in a planned application. This sort of collector offers sensible prospects for solar power use and it's appropriate for small- and large-scale applications. Some prototypes are tested. Its operation temperature was two hundred °C, its dimensions were sixteen m long × four m high and it consisted of eleven primary reflectors.

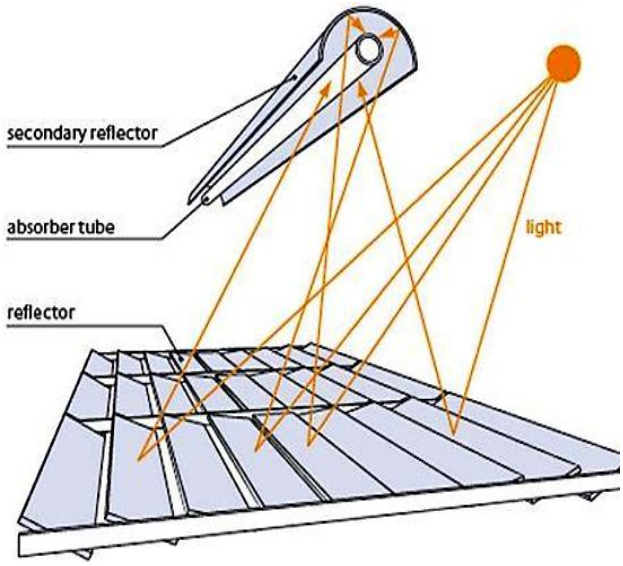


Figure 11: Linear Fresnel system [26]

2.6 Parabolic trough system

A distinction among this technique and hence a Liner Augustine Jean Fresnel system is parabolic trough system utilizes umbrella shaped formed reflectors. A concentration ratio will around eighty or additional .In absorber tube collected energy oh HT fluid will run. Parabolic trough technique provides energy at a temperature of up to 400°C. This energy is provided to run either a straightforward Jeannette Rankin cycle or hybrid system. A warmth transfer fluid that is employed to soak up warmth will be either water or artificial oils. Fig.12 depicts parabolic trough system components.

2.7 Parabolic trough solar power plants

Under this topic an operation eventualities &hence a totally different installation configuration for parabolic trough systems square measure described. Additionally, as this technique utilizes a sun following system, a used following system is in brief mentioned.

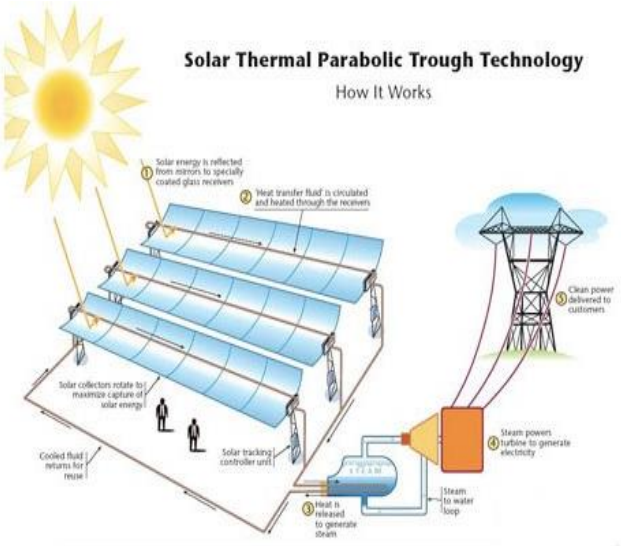


Figure 12: Parabolic trough system [27]

2.8 The sun chase system

Homeward-bound rotational axis, wherever collectors area unit. A system unceasingly drives collectors from east at sunrise to west at sunset. Little motors area unit accustomed drive this chase system. Fig.13 depicts solar furnace system concept. Since solely direct radiation is focused PTS utilizes a sun chase system for confirm most potency of a concentrating method

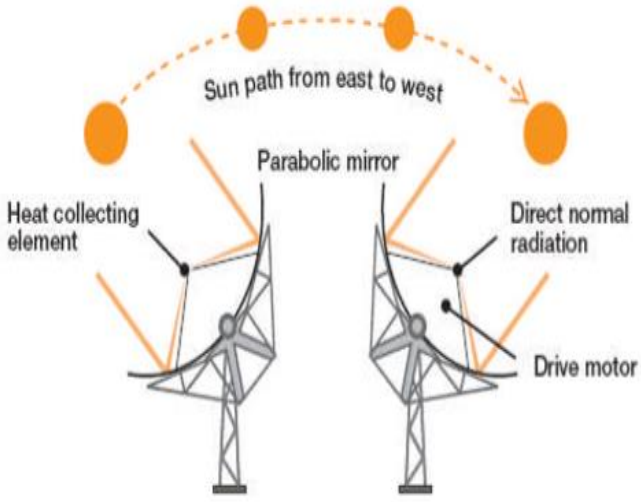


Figure 13: The sun chase system [28]

2.9 Solar desalinization

A potential of victimization star PTS in H_2O desalinization, wherever a star field is attached to water desalinization like multi-stage flash distillation (MSF) or multi impact

distillation. Below Fig. 14 depicts solar desalinization.

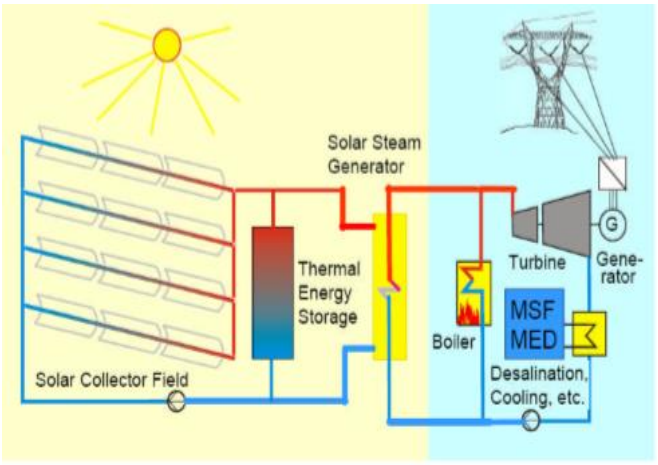


Figure 14: Solar desalinization [29]

2.10 Integrated solar combined cycle

An ISCC system may be a combination of a star field and gas turbine-combined cycle. A waste heat from a turbine is employed to come up with some steam to be distended during a turbine. Additionally, a star field provides further heat to a thermal cycle. An extra heat from a star field ends up in electricity generation increase throughout daylight time. This mixture ends up in up a thermal potency. A advantages of using this technology are to beat some issues associated with start-up and close up in alternative energy plants, scale back a cost of capital and improve a solar-to-electricity potency. Below Fig. 15 shows ISCC.

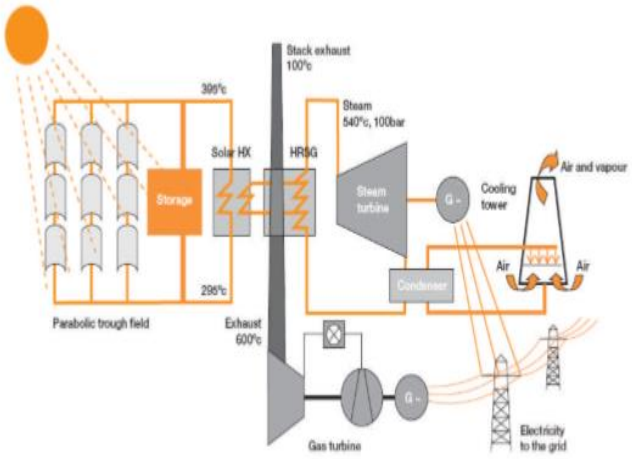


Figure 15: Integrated solar combined cycle [30]

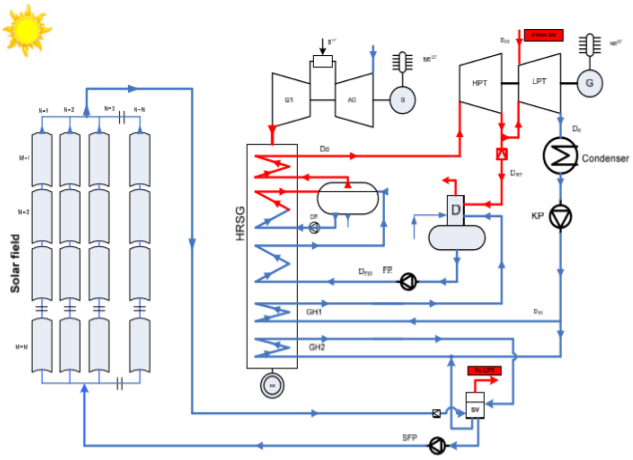


Figure 16: The proposed design for ISCC [31]

2.11 The proposed design

A planned style of a ISCC is shown in Fig.16. Its associate integration between a traditional combined cycle station (gas & steam turbine) and star field, supported a parabolic trough solar furnace. HRSG is one in every of a CC parts. It's wont to recover a warmth loss from a turbine exhaust gases. Most advanced electricity generation gas turbines area unit capable of being connected to heat recovery units. A most parts of a planned ISCC are: turbine unit, HRSG, turbine unit, and star field supported parabolic trough technology.

A turbine unit is that a major energy resource for a political leader cycle. HRSG that could be a device won't to recover heats from high temp. Gases streams (used with gas turbines). The HRSG consists 3 main sections, i.e. superheating section, evaporator & saver. The saver will pump up the feed water temp. To the saturation temperature to recover the max amount heat as potential from the gases stream. Then the steam generator (evaporator) converts the feed water to saturated steam at the HRSG drum's pressure. The superheating section will increase the steam temperature to the required temperature (HRSG, 2007). The projected style includes victimization 2 gas heaters. The primary warmer GH1 preheats water within the HRSG. Additionally an evaporator is employed within the HRSG to come up with some steam to produce some energy

3. Mathematical analysis of the integrated solar combined cycle

An approach wont to investigate various parts of a ISCC powerhouse is described. It includes a physics analysis of main objectives.

3.1 Combined cycle thermodynamics analysis

In an ISCC cycle a sole objective of analyzing a gas turbine unit. It is about to enumerate the energy's wastage in terms of gases at the exit. Under this section we will try to examine the performances of the cc that is combustion chamber,

performance of turbine, and last, compressor.
The equation for heater/heat exchanger of Brayton and Rankine cycle:

$Q_{hex} = \dot{m}_b(h_{4,b} - h_{1,b}) = \dot{m}_r(h_{1,r} - h_{4,r})$ (1)
The equation of the expander of Brayton cycle:
$n_{T,s} = h_2 - h_1 / h_3 - h_1$ (2)
$W_T = \eta_T*(h_1 - h_2)$ (3)
Where $\eta_T$is the mechanical efficiency of expander.
The equation of condenser of trans-critical cycle:
$Q_{cond} = \dot{m}(h_4 - h_1)$ (4)
The equation of fluid pump of Rankine cycle:
$W_{pump} = h_4 - h_3$ (5)
The work of turbine of trans-critical cycle.
$W_T = h_3 - h_4$ (6)
The work of turbine of Rankine cycle
$W_T = h_{1,r} - h_{2,r}$ (7)
The work of pump of trans-critical cycle.
$W_C = h_2 - h_1$ (8)
The work of pump of Rankine cycle
$W_P = h_{4,r} - h_{3,r}$ (9)
Equation for the net system output:
$W_{sys} = W_{NET,Rankine} + W_{NET,Brayton} + W_{NET,transcritical}$ (10)
Equation for finding the cycle thermal efficiency:
$\eta_{th} = W_{sys}/Q_{hex}$ (11)

3.2 Exergy Analysis

Economizer

$E_{rev,EC,b}=m_{,dot,b}*((h_{bv}-h_{5,b})-T_{amb}*(s_{bv}-s_{5,b}))$
$E_{rev,EC,r}=m_{dot,r}*((h_{4,r}-h_{rv})-T_{amb}*(s_{4r}-s_{rv}))$
$E_{rev,EC,total}=E_{rev,EC,b}+E_{rev,EC,r}$
(Total exergy in Economizer)

Heater

$E_{rev,HE,b}=m_{_dot_b}*((h_{bs}-h_{bv})-T_{amb}*(s_{bs}-s_{bv}))$
$E_{rev,HE,r}=m_{dot,r}*((h_{rv}-h_{rs})-T_{amb}*(s_{rv}-s_{rs}))$
$E_{rev,HE,total}=E_{rev,HE,b}+E_{rev,HE,r}$
(Total exergy in Heater)

Super heater

$E_{rev,SH,b}=m_{_dot,b}*((h_{4,b}-h_{bs})-T_{amb}*(s_{4,b}-s_{bs}))$
$E_{rev,SH,r}=m_{dot,r}*((h_{rs}-h_{1,r})-T_{amb}*(s_{rs}-s_{1,r}))$
$E_{rev,SH,total}=E_{rev,SH,b}+E_{rev,SH,r}$
(Total exergy in Super Heater)

Overall hex

$E_{rev,HEX,b}=m_{_dot,b}*((h_{4,b}-h_{5,b})-T_{amb}*(s_{4,b}-s_{5,b}))$
$E_{rev,HEX,r}=m_{dot,r}*((h_{4,r}-h_{1,r})-T_{amb}*(s_{4,r}-s_{1,r}))$

The exergy in Overall Heat Exchanger is given by following expression
$E_{rev,HEX},total=E_{rev,HEX,b}+E_{rev,HEX}$

4. Result and Discussion

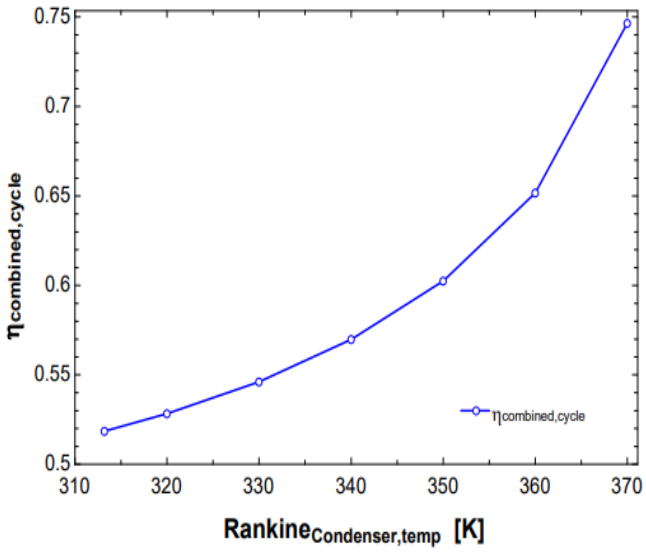


Figure 17: Variation between combined cycle efficiency and condenser temperature of Rankine Cycle

Fig. 17 shows the variation between combined cycle efficiency and condenser temperature of Rankine Cycle. It can be seen that on increasing the condenser temperature, efficiency is also increasing. The combined cycle efficiency is 56.98% at condenser temperature 340K.

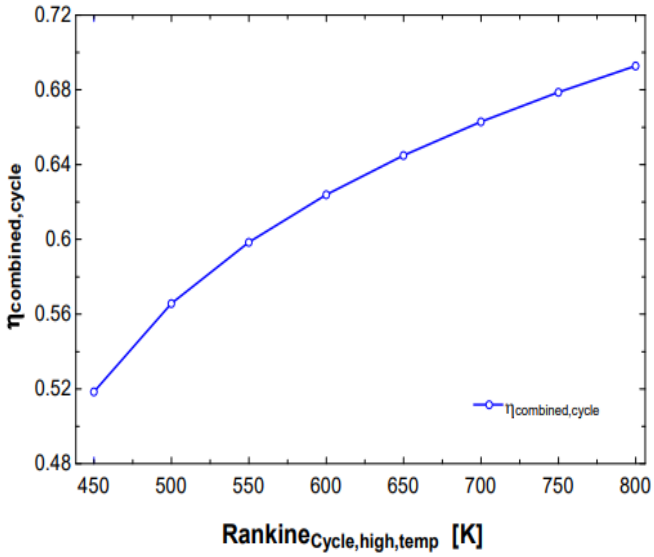


Figure 18: Variation between combined cycle efficiency and cycle high temperature of Rankine Cycle

Fig.18 shows the variation between combined cycle efficiency and cycle high temperature of Rankine Cycle. It can be seen that on increasing the cycle high temperature, efficiency is also increasing. The combined cycle efficiency is 51.85% at temperature 450K.

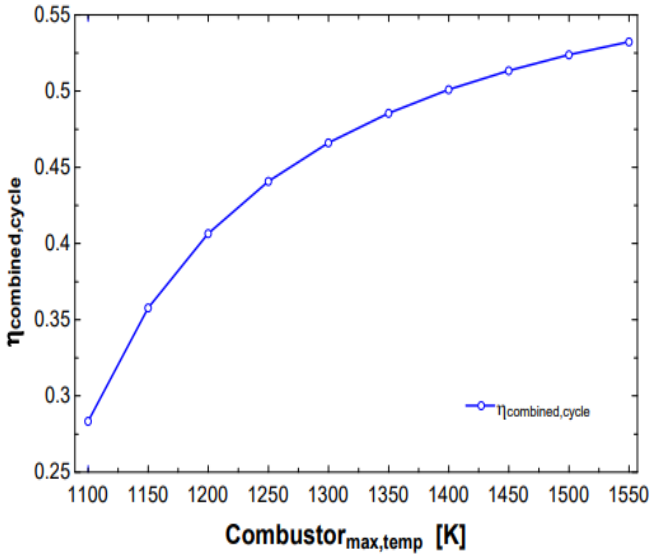


Figure 19: Variation between combined cycle efficiency and combustor maximum temperature of Brayton Cycle

Fig.19 shows the variation between combined cycle efficiency and combustor maximum temperature of Brayton Cycle. It can be seen that on increasing the combustor temperature, efficiency is also increasing. The combined cycle efficiency is 44.08% at combustor temperature 1250K

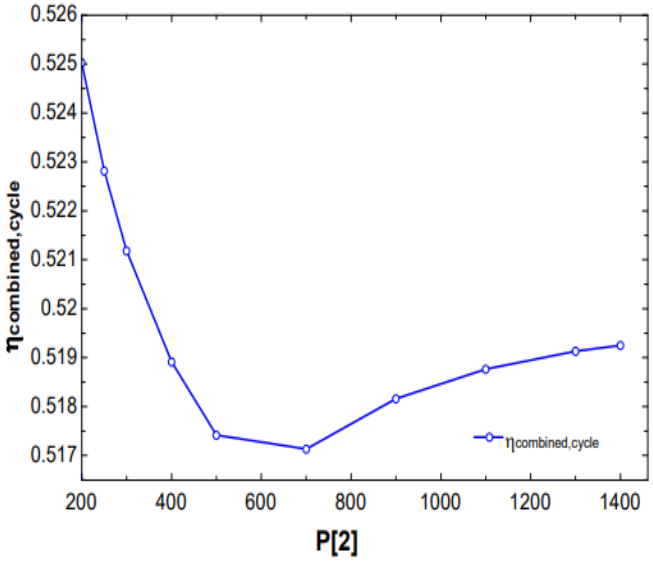


Figure 20: Variation between combined cycle efficiency and exit pressure of pump of trans-critical cycle.

Fig. 20 shows the variation between combined cycle efficiency and exit pressure of pump of trans-critical cycle. It can be seen that on increasing the pressure, efficiency is decreasing and after 700 kPa efficiency is slightly increasing. The combined cycle efficiency is 51.89% at pressure 400 kPa

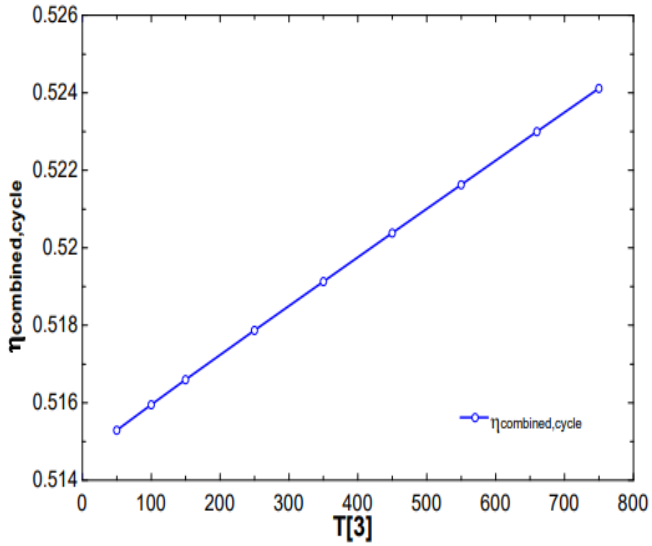


Figure 21: Variation between combined cycle efficiency and turbine inlet temperature of trans-critical Cycle.

Fig. 21 shows the variation between combined cycle efficiency and turbine inlet temperature of trans-critical Cycle. The can be seen that on increasing the temperature, efficiency is also increasing. The combined cycle efficiency is 51.79% at temperature 250ºC.

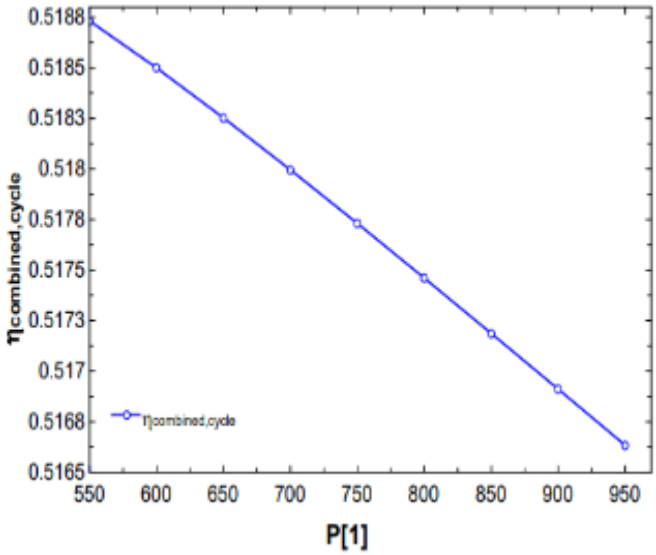


Figure 22: Variation between combined cycle efficiency and inlet pump pressure of trans-critical Cycle

Fig.22 shows the variation between combined cycle efficiency and inlet pump pressure of trans-critical Cycle. It can be seen that on increasing the pressure, efficiency is decreasing. The combined cycle efficiency is 51.8% at pressure 700 kPa.

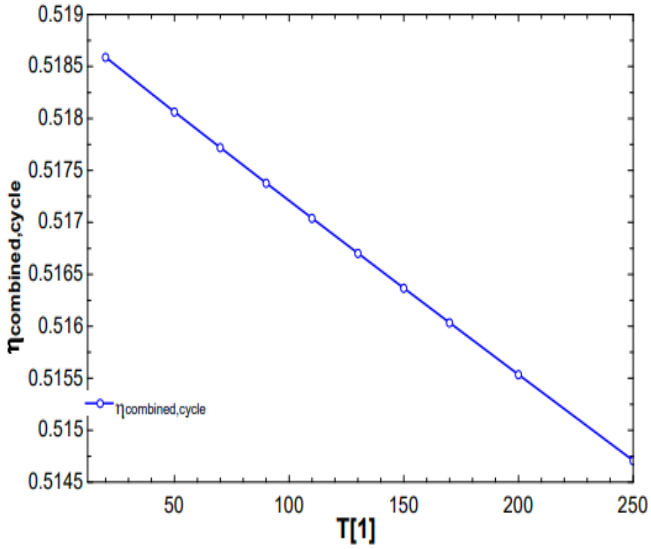


Figure 23: Variation between combined cycle efficiency and turbine inlet temperature of trans-critical Cycle

Fig.23 shows the variation between combined cycle efficiency and turbine inlet temperature of trans-critical Cycle. It can be seen that on increasing the temperature, efficiency is decreasing. The combined cycle efficiency is 51.7% at temperature 110ºC.

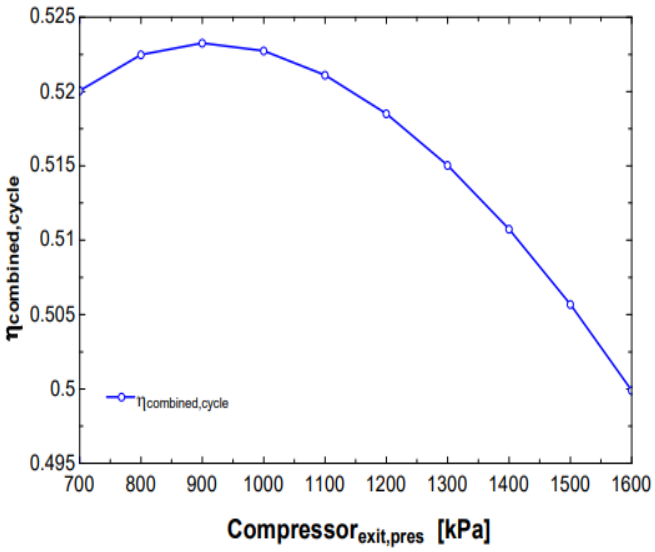


Figure 24: Variation between combined cycle efficiency and compressor exit pressure of Rankine Cycle

Fig. 24 shows the variation between combined cycle efficiency and compressor exit pressure of Rankine Cycle.

It can be seen that on increasing the pressure, efficiency is increasing up to 900 kPa and after that it keeps on decreasing. The combined cycle efficiency is 52.11% at compressor exit pressure 1100 kPa.

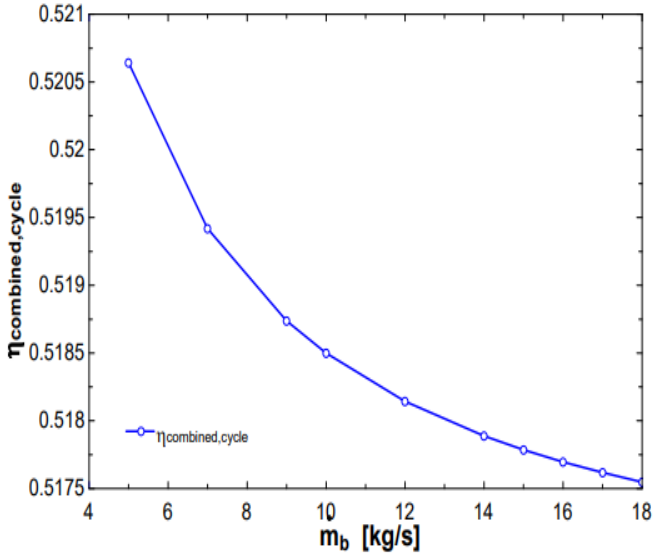


Figure 25: variation between combined cycle efficiency and mass flow rate of Brayton Cycle

Fig. 25 shows the variation between combined cycle efficiency and mass flow rate of Brayton Cycle. It can be seen that on increasing the mass flow rate, efficiency is decreasing. The combined cycle efficiency is 51.81% at mass flow rate 12 kg/sec.

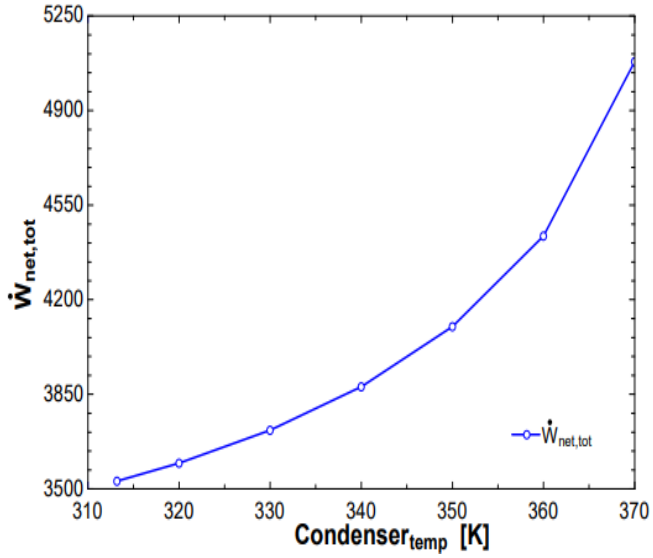


Figure 26: Variation between combined cycle net-work and condenser temperature of Rankine Cycle

Fig.26 shows the variation between combined cycle net-work and condenser temperature of Rankine Cycle. It can be seen that on increasing the condenser temperature, combined net-work is also increasing. The combined cycle net-work is 3878 kW at condenser temperature 340K.

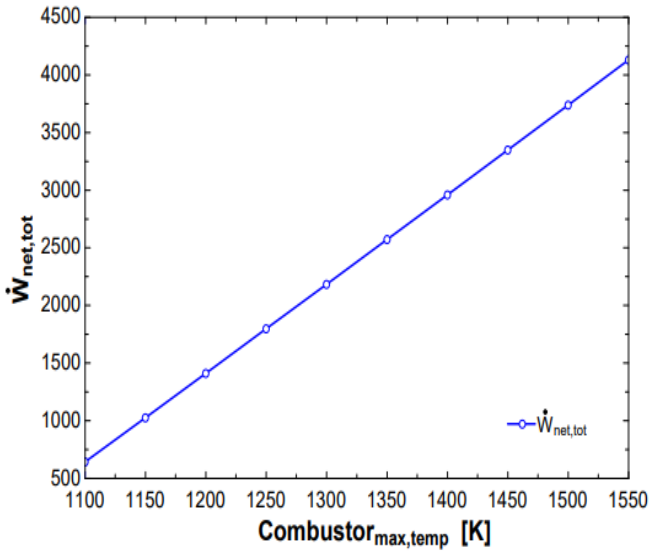


Figure 27: Variation between combined cycle net-work and cycle high temperature of Rankine Cycle

Fig.27 shows the variation between combined cycle net-work and cycle high temperature of Rankine Cycle. It can be seen that on increasing the cycle high temperature, net-work is also increasing. The combined cycle net-work is 3529 kW at temperature 450K.

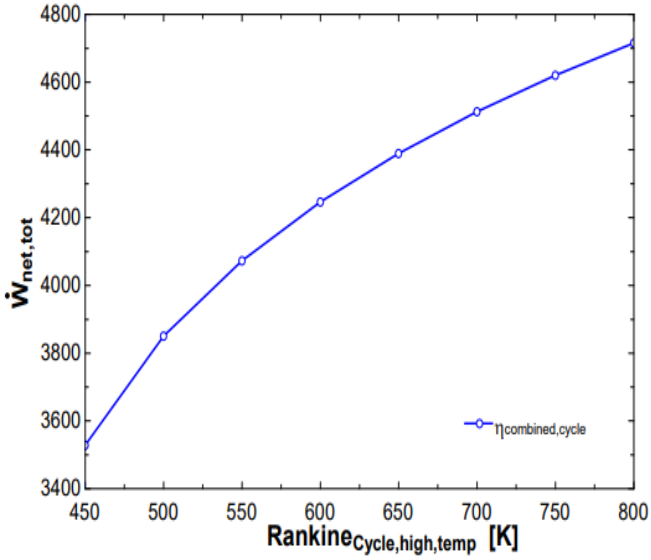


Figure 28: Variation between combined cycle net-work and combustor maximum temperature of Brayton Cycle

Fig. 28 shows the variation between combined cycle net-work and combustor maximum temperature of Brayton Cycle. It can be seen that on increasing the combustor temperature, net-work is also increasing. The combined cycle net-work is 1798 kW at combustor temperature 1250K.

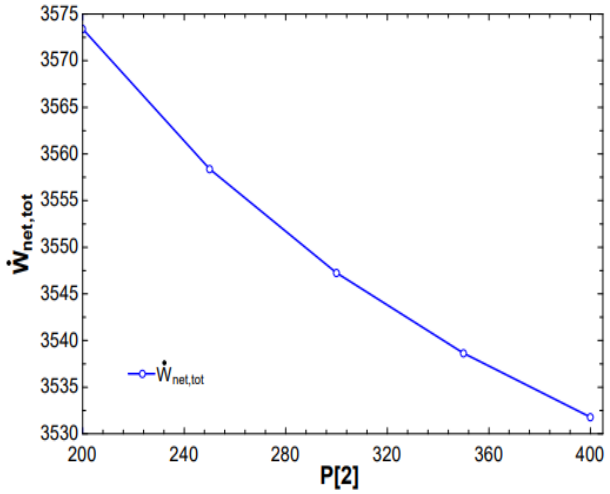


Figure 29: Variation between combined cycle net-work and exit pressure of pump of trans-critical cycle

Fig. 29 shows the variation between combined cycle net-work and exit pressure of pump of trans-critical cycle. It can be seen that on increasing the pressure, net-work is decreasing. The combined cycle net-work is 3532 kW at pressure 400 kPa.

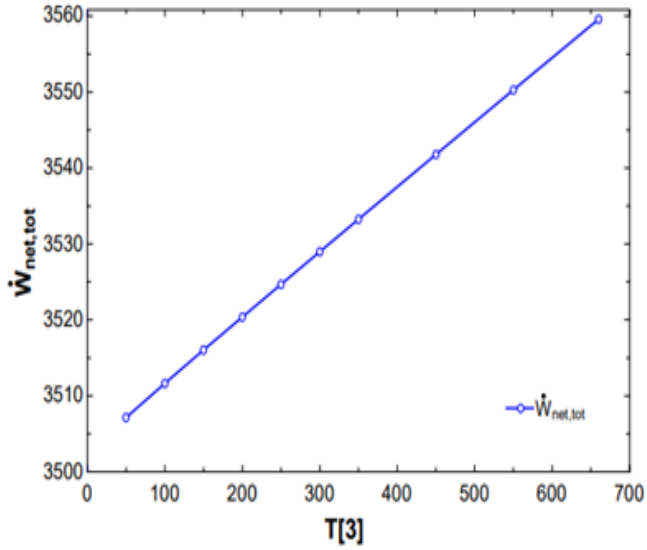


Figure 30: Variation between combined cycle net-work and turbine inlet temperature of trans-critical Cycle

Fig. 30 shows the variation between combined cycle net-work and turbine inlet temperature of trans-critical Cycle. It can be seen that on increasing the temperature, net-work is also increasing. The combined cycle net-work is 3525 kW at temperature 250ºC.

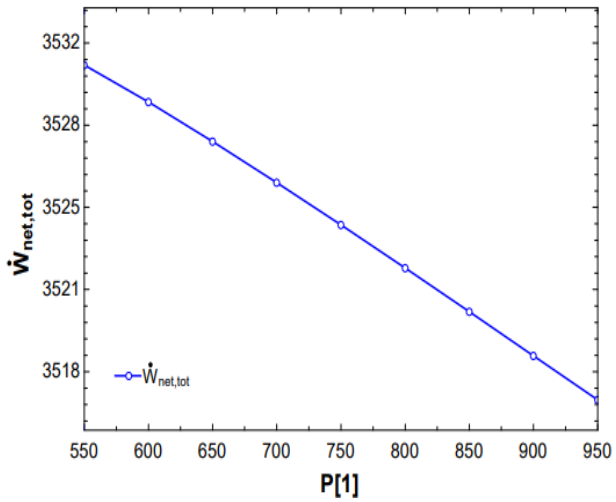


Figure 31: Variation between combined cycle net-work and inlet pump pressure of trans-critical Cycle

Fig.31 shows the variation between combined cycle net-work and inlet pump pressure of trans-critical Cycle. It can be seen that on increasing the pressure, net-work is decreasing. The combined cycle net-work is 3526 kW at pressure 700 kPa

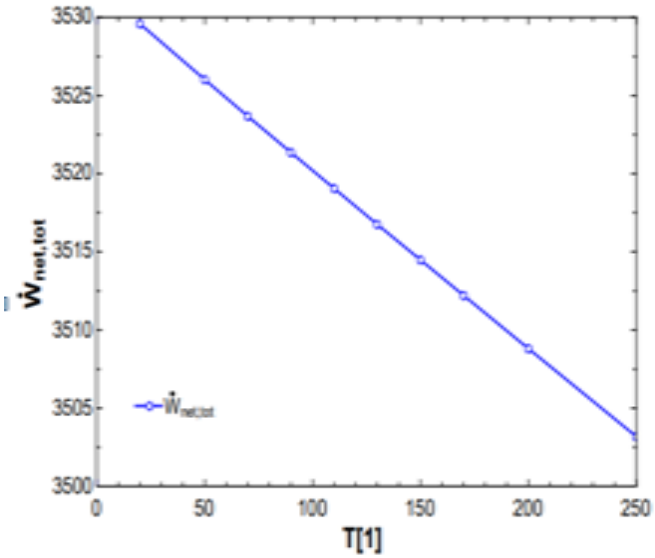


Figure 32: Variation between combined cycle net-work and turbine inlet temperature of trans-critical Cycle

Fig. 32 shows the variation between combined cycle net-work and turbine inlet temperature of trans-critical Cycle. It can be seen that on increasing the temperature, net-work is decreasing. The combined cycle net-work is 3519 kW at temperature 110ºC.

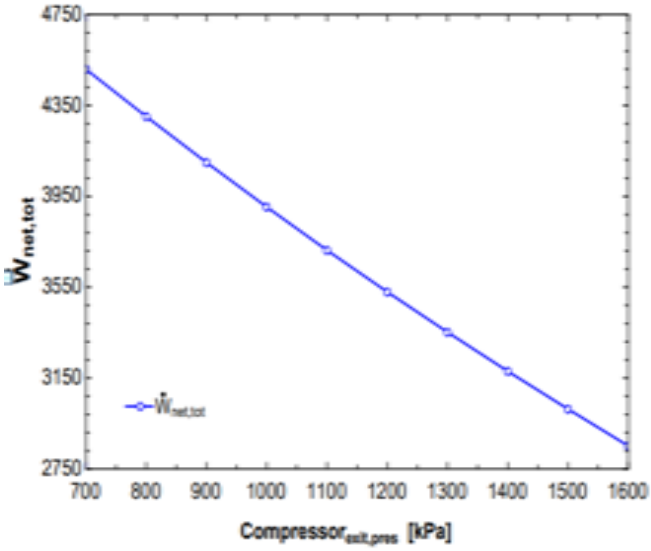


Figure 33: variation between combined cycle net-work and compressor exit pressure of Rankine Cycle

Fig.33 shows the variation between combined cycle net-work and compressor exit pressure of Rankine Cycle. It can be seen that on increasing the pressure, net-work is decreasing. The combined cycle net-work is 3713 kW at compressor exit pressure 1100 kPa.

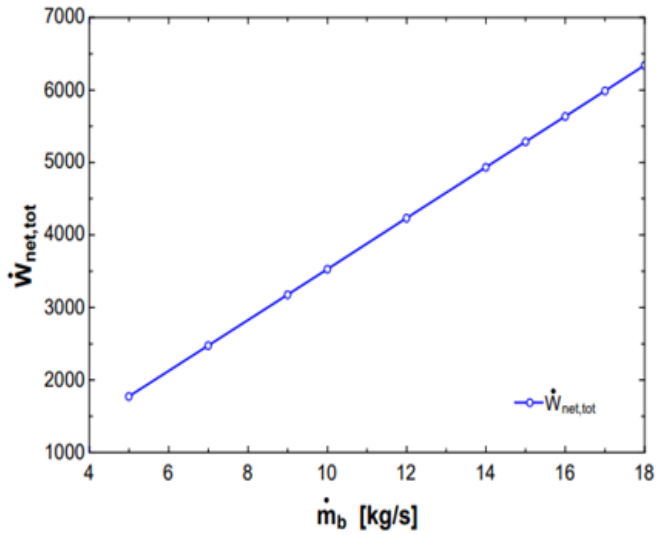


Figure 34: Variation between combined cycle net-work and mass flow rate of Brayton Cycle

Fig. 34 shows the variation between combined cycle net-work and mass flow rate of Brayton Cycle. It can be seen that on increasing the mass flow rate, efficiency is increasing. The combined cycle net-work is 4232 kW at mass flow rate 12 kg/sec

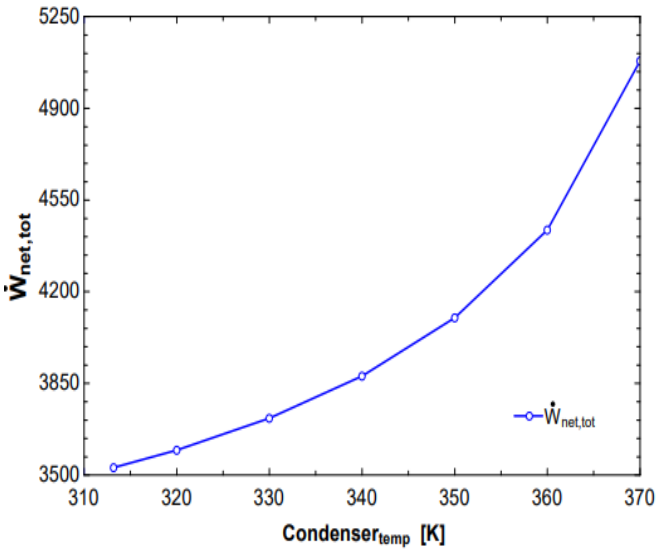


Figure 35: Variation between combined cycle net-work and condenser temperature of Rankine Cycle

Fig.35 shows the variation between combined cycle net-work and condenser temperature of Rankine Cycle. It can be seen that on increasing the condenser temperature, combined net-work is also increasing. From Table 19 combined cycle net-work is 3878 kW at condenser temperature 340K.

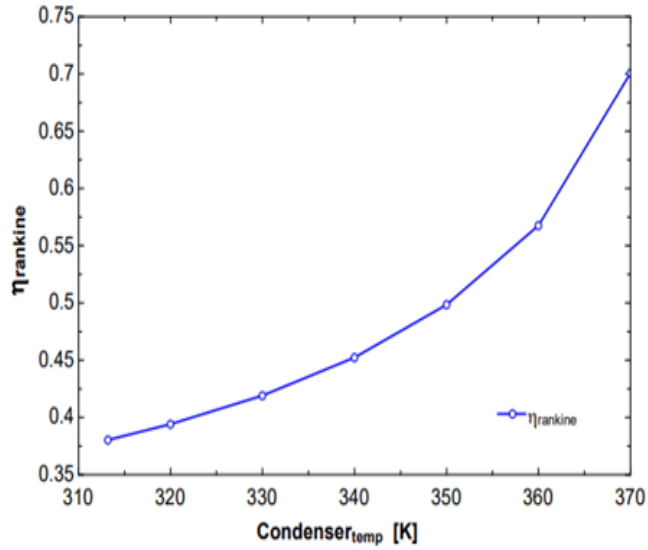
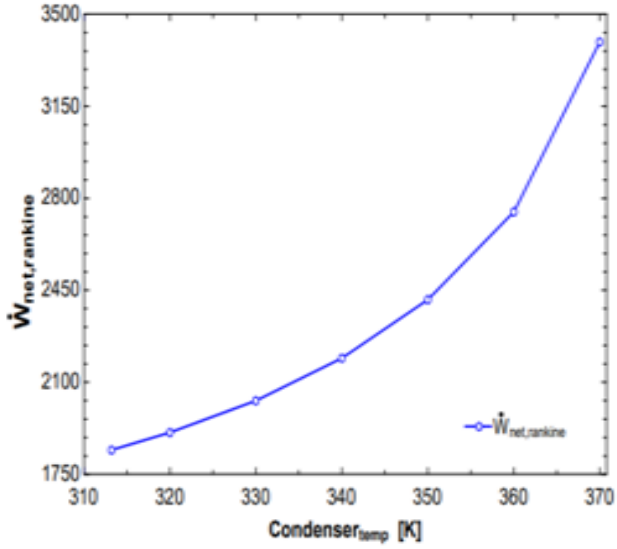


Figure 36: Variation between Rankine cycle efficiency and condenser temperature

Fig.36 shows the variation between Rankine cycle efficiency and condenser temperature. It can be seen that on increasing the condenser temperature, efficiency is also increasing. The Rankine cycle efficiency is 45.25% at condenser temperature 340K



Fig*ure 37: Variation between Rankine cycle net-work and condenser temperature.*

Fig. 37 shows the variation between Rankine cycle net-work and condenser temperature. It can be seen that on increasing the condenser temperature, Rankine net-work is also increasing. The combined cycle net-work is computed as 2191 kW at condenser temperature 340K.

5. Conclusion

In this study, carbon dioxide cycles is also investigated for recovering the low grade heat. In this thesis, three cycle are combined i.e. Brayton, Rankine and Trans critical cycle in which fluid taken are Ideal gas (air), Steam and Carbon dioxide respectively. The mass flow rate of Brayton cycle is 10 kg/sec and atmospheric temperature is 25ºC.
The following conclusions were drawn from thermodynamic analysis of above mentioned cycles have been shown below.

- On increasing condenser temperature of Rankine cycle, combined efficiency of cycle and combined net-work both are increasing.
- On increasing the mass flow rate of Brayton cycle, combined efficiency of cycle is decreasing and combined net-work is increasing.
- On increasing the exit pressure of pump in trans-critical cycle, combined net work is decreasing also combined cycle efficiency is decreasing.
- On increasing the inlet temperature of turbine of trans-critical cycle, combined cycle efficiency and combined net-work is increasing.
- On increasing the inlet pressure of pump of trans-critical cycle, combined cycle efficiency and combined net work both are increasing.

References

[1] Al-Hamdan, Q. and Ebaid, M. (2006), "Modeling and Simulation of a Gas Turbine Engine for Power Generation", vol. 128, no. 2, p. 311.

[2] Alrobaei, H. (2006a), Integrated Gas Turbine - Solar Power Plant, available at: http://www.energycentral.com/centers/knowledge/whitepapers/report.cfm?rid=102295 (accessed 20/11/2006)

[3] Alrobaei, H. (2006b), Hybrid Solar/Fossil Fuel Cogeneration Power Plant, available at: http://www.energycentral.com/centers/knowledge/Hybrid Solar/Fossil Fuel Cogeneration Power Plant/ (accessed 22/07/2007).

[4] Alrobaei, H. (2004), Effectiveness of Semi-Combined Solar Power Plant, available at: http://www.softinway.com/science/articles (Accessed 23/07/2007).

[5] Alrobaei, H. (1998), "Thermodynamic Parametric Study for Combined Cogeneration Power Plants", Journal of Engineering, vol. 38, pp. 8091.

[6] ALSTOM (2007), Product Brochure GT12E2 & GT8C2, available at: http://www.power.alstom.com/home/elib/ (accessed 30/07/2007).

[7] ANU (2007), (The Australian National University), available at: http://solar.anu.edu.au (accessed 11/05/2007).

[8] ASHRAE (2003), Applications Handbook (SI), ASHRAE, Atlanta, GA, USA.

[9] Badran, O. and Eck, M. (2006), "The application of parabolic trough technology under Jordanian climate", Renewable Energy, vol. 31, no. 6, pp. 791-802.

[10] Beasley, O. (1994), "Induced Draft Fan Innovation for Heat Recovery Steam Generators", Journal of Engineering for Gas Turbines and Power, vol. 116, pp. 402-405.

[11] Becker, M. and Trieb, F. (2000), Solar Thermal Power Plants, available at: http://www.solarpaces.org/Library/library.htm

[12] Bernardes, D. and Weinrebe, A. (2003/12), "Thermal and technical analyses of solar chimneys", Solar Energy, vol. 75, no. 6, pp. 511-524.

[13] CSES, (2007), Monthly Daily Meteorological Data, Center for Solar Energy Studies – Libya, Tripoli.

[14] Nakata, T., Hasegawa, T., Ninomiya, T. and Sato, M. (1997), Development of a 1,500°C-Class Gas Turbine Combustor for IGCC, Central Research Institute of Electric Power Industry.

[15] DLR, Institute of Technical Thermodynamics and Section Systems Analysis and Technology Assessment (2006a), Trans-Mediterranean Interconnection for Concentrating Solar Power, , Federal Ministry for the Environment; Nature Conservation and Nuclear Safety Germany, Stuttgart, Germany.

[16] DLR (2006b), Concentrating Solar Power for the Mediterranean Region, available at: http://www.dlr.de/tt/med-csp (accessed 11/11).

[17] DLR (2002), Concentrating Solar Power Now, D-11055, The Federal Ministry for the Environment, Nature Conservation and Nuclear Safety (BMU), Berlin, Germany.

[18] Duffie, J. (1991), Solar Engineering of Thermal Processes, 2nd Ed, Wiley, New York.

[19] Eastop, T. and McConkey, A. (1993), Applied thermodynamics for engineering technologists, 5th ed, Longman, London.

[20] Eck, M. and Hirsch, T. (2007/2), "Dynamics and control of parabolic trough collector loops with direct steam generation", Solar Energy, vol. 81, no. 2, pp. 268-279.

[21] Elgady, A. (2007), Data about the Libyan power sector, GECOL (the General Electricity Company Of Libya), Zawia, Libya.

[22] El-Osta, W. (2003), "Prospects of wind power plants in Libya: a case study", Renewable Energy, vol. 28, no. 3, pp. 363-371.

[23] ETB (2007), Gases and Densities, Engineering Tool Book ,available at: http://www.engineeringtoolbox.com (accessed 17/08/2007.

[24] European Commission (2004), European Research on Concentrated Solar Thermal Energy, Brussels.

[25] Flagso GmbH (2007), Brief Technical Description of the SKAL-ET Collector, available at: http://www.flagsol.com (accessed 11/08/2006).

[26] Flynn, D. (2003), Thermal power plant simulation and control, 1s Ed, Institution of Electrical Engineers, London, UK.

[27] GECOL (2006), GECOL Annual Report 2006, , General Electricity Company of Libya, Tripoli, Libya.

[28] GECOL (2005), GECOL Annual Report 2005, General Electricity Company of Libya, Tripoli, Libya.

[29] Gill, A. (1984), Power Plant Performance, 1st Ed, Butterworths, London.

[30] Greenpeace International (2003), Solar Thermal Power 2020 Exploiting the Heat from the Sun to Combat Climate Change, ISBN:9073361-82-6, Greenpeace and the European Solar Thermal Industry Association (2003).

[31] Herrmann, U., Kelly, B. and Price, H. (2004/0), "Two-tank molten salt storage for parabolic trough solar power plants", Energy, vol. 29, no. 56, pp. 883-893.

[32] Hosseini, R., Soltani, M. and Valizadeh, G. (2005/8), "Technical and economic assessment of the integrated solar combined cycle power plants in Iran", Renewable Energy, vol. 30, no. 10, pp. 1541-1555.

[33] Hottel, H. C. (1976), "A simple model for estimating the transmittance of direct solar radiation through clear atmospheres", Solar Energy, vol. 18, no. 2, pp. 129-134.

[34] HRSG User's group (2007), HRSG User's Handbook, available at: http://www.hrsgusers.org/practices.php (accessed 01/08/2007).

[35] Huang, M. and Gramoll, K. (2007), Actual Gas-turbine Cycle, available at: https://ecourses.ou.edu/home.htm (accessed 25/07/2007).

[36] Incropera, F. (2002), Fundamentals of Heat and Mass Transfer, 5th ed, John Wiley & Sons, New York.

[37] IWAI, M. (2003), "Thermodynamic Table for Performance Calculations in Gas Turbine Engine", Proceedings of the International Gas Turbine Congress, 2-7 November 2003, Tokyo.

[38] Jacobson, E., Ketjoy, N., Nathakaranakule, S. and Rakwichian, W. (2006), "Solar parabolic trough simulation and application for a hybrid power plant in Thailand", vol. 32, no. 2, pp. 187-199.

[39] Jones, J. (2007a), "CSP lifts off. Nevada Solar One comes to life ", Renewable Energy World, vol. 10, no. 3, pp. 36-42.

[40] Jones, J. (2007b), "Solar Movement", Renewable Energy World, vol. 10, no. 4, pp. 168-172.

[41] Lane, D. (2007), Brayton Cycle: The Ideal Cycle for Gas-Turbine Engines In Relation to Power Plants, available at: http://coeweb.engr.unr.edu/me/ (accessed 22/07/2007).

[42] Mills, D. (2004), "Advances in solar thermal electricity technology", Solar Energy, vol. 76, no. 1-3, pp. 19-31.

[43] Mustafa, E. (2007), Monthly Daily Meteorological Data, LREC Libyan Renewable Energy Centre.

[44] Najjar, Y. (1996/2), "Enhancement of performance of gas turbine engines by inlet air cooling and cogeneration system", Applied Thermal Engineering, vol. 16, no. 2, pp. 163-173.

[45] Natan, S., Barnea, D. and Taitel, Y. (2003/11), "Direct steam generation in parallel pipes", International Journal of Multiphase Flow, vol. 29, no. 11, pp. 1669-1683.

[46] NREL, (2007), U.S. Parabolic Trough Power Plant Data, available at: http://www.nrel.gov/csp/troughnet/power_plant_data.html

Nomenclature

A	Altitude [m]
Ap	Total outer area of the receiver tube [m^2]
ASF	Total solar field aperture area
[m^2]	The specific fuel consumption of the gas turbine unit [tonne/MWh]
BGT	Gas turbine fuel consumption [tonne/h]
C	Solar collector concentration ratio [-]
Cp	Specific heat [kJ/kg.K]
DB	Fuel saving [tonne/h]
Dci	Cover inner diameter [m]
Dco	Cover outer diameter [m]
Devap	Steam mass loss from the deaerator [kg/s]
DFW	Mass flow of feed water [kg/s]
DK	Water mass flow in plant condenser [kg/s]
DLoss	Steam loss [kg/s]
Do	Steam mass flow at the turbine inlet (reference point) [kg/s]
DRK	Water mass flow in GH1 [kg/s]

DRK2	Water mass flow in GH2 [kg/s]
DRT	The extracted steam to operate the plant deaerator [kg/s]
DSS	The generated stem due to solar field Contribution [kg/s]
Dti	Receiver inner diameter [m]
Dto	Receiver outer diameter [m]
h'	Saturated water specific enthalpy [kJ/kg]
h"	Saturated steam specific enthalpy [kJ/kg]
Ib	Beam solar radiation [W/m^2]
Id	Diffuse solar radiation [W/m^2]
Isc	Solar constant [W/m^2]
Iso	Extra-terrestrial solar radiation [W/m^2]
K	Receiver thermal conductivity [W/m. °K]
kc	Cover thermal conductivity [W/m. °K]
Ke	Cover extinction coefficient [m-1]
l	Collector length [m]
N	Number of collectors in each row [-].
M field [kg/s] m.	Water mass flow for each row in the ^{solar}
SF [kg/s]	Water mass flow for whole solar field
mC gas turbine unit	Relative air mass flow for blades cooling in
mf unit [kg fuel/kg air]	The relative fuel mass flow for gas turbine
mgas [kg/s] mK	Gases mass exhaust from gas turbine unit
N	Number of rows of solar field [-]
n2	Cover refractive index [-]
NEGT	Gas turbine output [MW]
NEST	Steam turbine output [MW]
NFP [MW]	Energy consumption by water feed pump
P	Pressure [bar]
PD	Deaerator pressure [bar]
PDE	Deaerator's evaporator pressure [bar]
Pk	Condenser pressure [bar]
PLPT	Pressure at LPT inlet [bar]
PLPTO	Pressure at LPT inlet for combined cycle operation [bar]
PSOSF	Design outlet pressure for solar field [bar]
Qc.v	Fuel calorific value [kJ/kg]
QL	Heat loss from solar collector [kW]
QSC	Useful heat from solar field [kW]
QSF	Nominal solar field output [kW]
Qu row) [kW]	Useful heat gain in solar field (for each
R	Gas constant [kJ/kg.°K]
S	Specific entropy [kJ/kg.°K]
Sb [W/m^2.°K]	Absorbed solar energy by receiver tube
T	Temperature [°C]
Ta	Ambient temperature [°C]
Tbw blades [°C]	The average temperature of gas turbine
Tex	Exhaust Gases temperature after HRSG °C]
Tfi	Water temperature at solar field inlet [°C]
Tfo	Water temperature at solar field outlet [°C]
TL	Disposed water temperature [°C]
TRFW1	Re-feed water temperature [°C]
TS [°C]	Temperature for ideal process (isentropic)
Ua	Wind Velocity [m/s]
UL	Solar collector loss coefficient [W/m^2.°K]
W	Collector aperture width [m]
Wa [kJ/kg]	Specific work done by gases in the GT
Wco GT [kJ/kg]	Specific work done by cooling air in the
We	Specific work for gas turbine unit [kJ/kg]
WK	Compressor specific work [kJ/kg]
WT	Total specific work of GT (gases + air) [kJ/kg]
XK	Steam/water dryness factor [%]

Abbreviations

AC	Air compressor
ANU	The Australian National University
CC	Combined cycle
CHP	Combined Heat and Power
CSES	Centre for Solar Energy Studies – Libya
CSP	Concentrating Solar Power
DE	The evaporator of the desecrator
DISS	Direct solar steam European project
DLR	German Aerospace Centre
DSG	Direct steam Generation
EC	European commission
ECC	Equivalent combined cycle
ETB	Engineering tool Book
EU-MENA	Europe, Mediterranean North African region
FP	Feed water pump
FV	Flash vessel
G	Electricity generator
GCC	Gas Turbine Combustion chamber
GECOL	General Electricity Company of Libya
GH	Gas heater
GT	Gas turbine
GTU	Gas turbine unit
HPT	High pressure turbine
HRSG	Heat Recovery Steam Generation
HTF	Heat Transfer Fluid
HVDC	High Voltage Direct Current
ISCC	Integrated Solar Combined Cycle Power Plant
LPT	Low pressure turbine
LREC	Libyan Renewable Energy Centre
MED	Multi Effect Desalination Unit
MSF	Multi Stage Flash Desalination Unit
NREL	National Renewable Energy Laboratory
RFWH	Re-feed water heater
SEEN	The Sustainable Energy and Economy Network
SEGS	Solar Electricity Generating Station
STU	Steam turbine unit
SV	Separator vessel
TRANS-CSP	Trans-Mediterranean Interconnection for Concentrating Solar Power
TREC	Trans-Mediterranean Renewable Energy Cooperation

Journal of Thermal Engineering, Vol. 5, No. 4, pp. 222-236, July, 2019
Yildiz Technical University Press, Istanbul, Turkey

THERMODYNAMIC PERFORMANCE ANALYSIS OF DEDICATED MECHANICALLY SUBCOOLED VAPOUR COMPRESSION REFRIGERATION SYSTEM

S. Agarwal[1], A. Arora[1,*], B. B. Arora[1]

ABSTRACT

In this work, the thermodynamic analysis of dedicated mechanically subcooled vapour compression refrigeration system is presented. A software based computer program in EES has been formulated for computation of the performance results. The effect of varation of degree of subcooling (5-30^{o}C), evaporator temperature (-20 to 10^{o}C) and condenser temperature (30-50^{o}C) has been investigated for energetic and exergetic performance of the system. The analysis of the system has been carried out using zero ODP and very low GWP (1 to 4) refrigerants viz.HFO-R1234ze and R1234yf to compare the performance of HFC-R134a. The results depicts that the COP and exergetic efficiency of dedicated subcooled VCR cycle are better than that of simple VCR cycle. Refrigerant R1234ze performs better than R1234yf and comparable to R134a.

Keywords: *VCR, Subcooling, COP, R1234ze, R1234yf, Exergetic Efficiency, Dedicated, R134a*

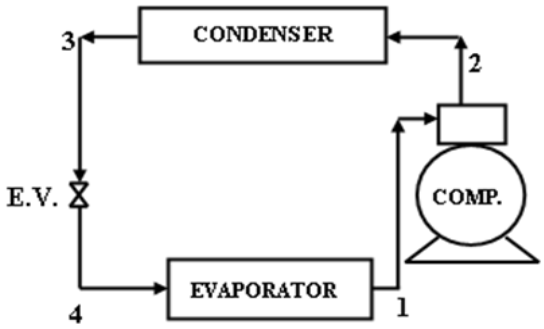
INTRODUCTION

Demand of energy, economy and eco-sustainability are the major considerations of a refrigeration system design. The demand of high grade energy is growing continuously in the developing countries viz. India. The cost of electricity is also increasing with its demand. The distortion of surrounding ecosystem with the use of high GWP refrigerants (HFCs) in the refrigeration and air-conditioning systems is a current issue. The high GWP HFCs have to be phased out as per the European Union (EU) regulation (Regulation (EU) No 517/2014 [1]). Many researchers have proved that the mechanically sub-cooling of vapour compression refrigeration (VCR) cycle improves its performance. Zubair [2, 3] and Zubair et al. [4] have investigated a dedicated and an integrated mechanical sub-cooling system from the thermodynamic standpoint. In these studies, it was found that the system performance had improved when operating in situations where the difference between the condensing and evaporating temperatures is large. Couvillion et al. [5] developed a mathematical model of a dedicated mechanical sub-cooling system by considering the individual component models of the equipment involved in the system. They found an improvement of 6-8% in the coefficient of performance (COP) and 20-170% in the capacity over a conventional simple cycle. Qureshi et al. [6] analysed that the load carrying capacity of the evaporator increased by approximately 0.5kW when R22 was subcooled in the main cycle by 5-8^{0}C. They noted that using the sub-cooling, the second-law efficiency of the cycle increased by an average 21% and this percentage increase is direcltly proportional to the ambient temperature. Khan and Zubair [7] have investigated finite time thermodynamic models of simple vapour compression refrigeration, two stage VCR system (TSS), dedicated mechanical sub-cool VCR system (DMSS), integrated mechanical sub-cooling VCR system (IMSS). They predicted that the optimum value of COP is obtained by varying the parameters of heat exchanger and reduced sub-cooler saturation temperature. Thornton et al. [8] have used dedicated mechanical sub-cooling VCR cycle to sub-cool the condenser exit of main VCR cycle. They predicted the optimum value of sub-cooling evaporator temperature which established the design rule for optimum distribution of heat exchanger area. Arora et al. [9, 10, 11], Arora and Kaushik [12] and Arora [13] have analysed vapour compression and absorption refrigeration systems for energetic and exergetic performance improvement. Llopis et al. [14, 15] have theoretically analysed the energy performance enhancement of CO_2 transcritical refrigeration system using dedicated mechanical sub-cooling cycle. They observed 20% increase in COP and 28.8% increases in cooling capacity of combined and reviewed sub-cooling method viz. internal heat exchanger, economizers, thermoelectric systems, dedicated sub-cooling methods of CO_2 refrigeration cycle. They concluded that the sub-cooling is a worth method to increase the performance of refrigeration cycle and deserve future development. Many researchers have suggested the replacement of R134a with R1234ze and R1234yf in refrigeration

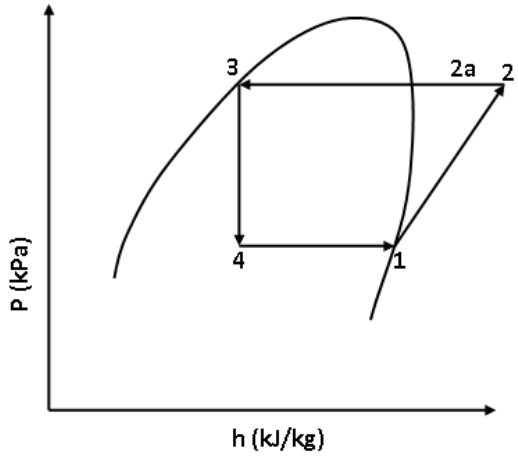
This paper was recommended for publication in revised form by Regional Editor Sandip Kale
[1]Department of Mechanical Engineering, Delhi Technological University, Delhi-110042, India.
**E-mail address: profakhilesharora@gmail.com*
Orcid id: 0000-0002-2363-7296, 0000-0002-5984-406X, 0000-0001-8022-5573
Manuscript Received 9 August 2017, Accepted 11 November 2017

systems. Mota-Babiloni et al. [16] reviewed the recent developments in commercial refrigeration and focused on system modifications (as dedicated sub-cooling or the implementation of ejectors), tri-generation technologies (electrical, heating and cooling demand) and better evaporation conditions control. They concluded that the hydrocarbon, HFO and CO_2 are HFC replacements. Kalla et al. [17, 18] reviewed alternative refrigerants for vapour compression refrigeration system and investigated the performance of R22 and its substitutes in air-conditioners. Yataganbaba et al. [19] performed exergy analysis on a two evaporator vapour compression refrigeration system using R1234yf, R1234ze and R134a as refrigerants. The two refrigerants R1234yf and R1234ze were good alternatives to R134a regarding their environment friendly properties. Dixit et al. [24-26] carried out energy and exergy analysis of absorption-compression cascade and waste heat driven triple effect refrigeration cycles.

The literature survey gathered from different established data sources viz. Google Scholar, Research gate, Science direct etc. reveals that the extensive research have been done on simple VCR cycle. Out of these, some studies based on dedicated mechanical sub-cool cycle have been considered to enhance the performance of the cycle. It is observed that limited work has been carried out to explore the performance of dedicated mechanical subcool cycle using HFO refrigerants viz. R1234yf and R1234ze.



(a)



(b)

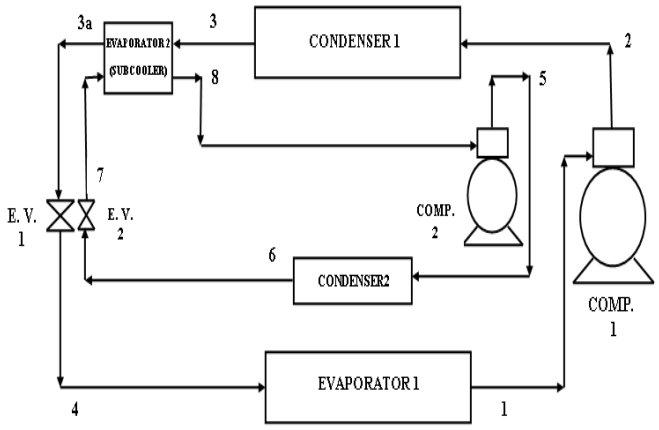
Figure 1. (a) Saturated Vapour Compression Refrigeration Cycle, (b) P-h diagram of saturated vapour compression refrigeration cycle

The present work investigates the thermodynamic performance of dedicated mechanically subcooled VCR cycle using R1234yf, R1234ze and R134a. The comparison of considered refrigerant has also been explored for drop-in replacement of R134a to R1234ze and R1234yf. The first and second law based performance parameters viz. COP and exergetic efficiency have been computed with the variation in degree of sub-cooling (DOS), evaporator temperature and condenser temperature.

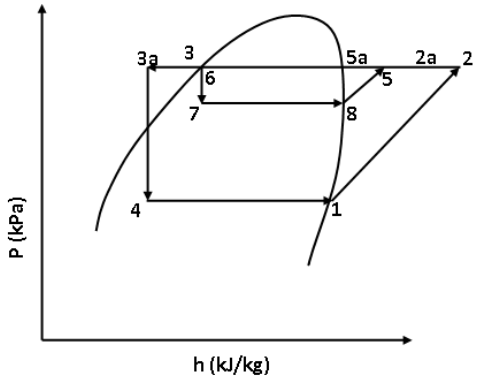
DESCRIPTION OF SYSTEM AND MODELLING

Description of System

The current work presents first and second law based thermodynamic analysis of dedicated mechanically subcooled vapour compression refrigeration system (DSC) using R1234ze, R1234yf and R134a. The mechanical sub-cooling of main VCR cycle has been carried out through the subcooler or evaporator of another simple VCR cycle designated as subcooler VCR cycle.



(a)



(b)

Figure 2. (a) Schematic diagram of dedicated mechanically subcooled vapour compression refrigeration cycle, (b) P-h diagram of dedicated sub-cooled vapour compression cycle

Figure 2(a) and 2(b) illustrate the schematic and P-h diagrams of dedicated mechanically subcooled vapour compression refrigeration system respectively. It comprises of two sections i.e. subcooled and subcooler vapour compression refrigeration cycles. The evaporator of subcooler VCR cycle receives the heat rejected by the condensate liquid refrigerant of subcooled VCR cycle and acts as subcooler. In this way, the two cycles are coupled together through subcooler or evaporator 2 in order to achieve low temperature in evaporator1. The refrigerants considered for the energy and exergy analysis of the system are R134a, R1234yf and R1234ze. The same refrigerant has been considered in the two sections simultaneously.

The subcooled VCR cycle is in lower section which consists of compressor1, evaporator1, expansion valve1, subcooler and condenser1 and the subcooler VCR cycle is upper section which consists of compressor2, condenser2, expansion valve2 and evaporator2 or subcooler. The vapour refrigerant leaving the evaporator1 of subcooled VCR cycle at state point1 enters into the compressor1 and is being compressed. The refrigerant at high temperature and pressure leaving the compressor1 at state point 2 and enters in to the condenser1. The liquid condensate has been subcooled from state point 3-3a by the evaporator2 of the subcooler cycle. The subcooled liquid refrigerant is throttled by the expansion valve1 from state point 3a-4. The low temperature liquid-vapour mixture of refrigerant enters in to the evaporator1 and low temperature is achieved in the evaporator1.

Similarly, the saturated vapour of the refrigerant leaving the evaporator2 or subcooler at state point 5 enters in to the compressor2. The compressed high pressure, high temperature vapour refrigerant enters into the condenser2 at state point 6. The vapour phase of the refrigerant is converted in to liquid in the condenser2. The liquid condensate refrigerant leaving the condenser2 at state point7 is expanded freely by the expansion valve2 from state 7-8. The low temperature liquid-vapour mixture of refrigerant enters into the evaporator2 or subcooler and absorbs heat from the condensate of condenser1 from state 8-5 and transformed into saturated vapour.

In this way, the mechanical sub-cooling is produced by the subcooler VCR cycle in the subcooler from states 3-3a. The evaporator temperature ($T_{e,SC}$) of the subcooler cycle is more than the evaporator temperature of subcooled cycle. The sub-cooling of liquid condensate refrigerant of subcooled cycle enhances the net specific refrigerating effect as shown in Figure 2 from states 4-1. The increase in refrigerating effect enhances the coefficient of performance (COP) of the subcooled cycle. However, the compressor work of the subcooler cycle reduces the COP of the overall cycle.

Thermodynamic Modelling

In the present work, the energy and exery analysis of dedicated mechanically subcooled vapour compression cycle have been carried out. In which the subcooler VCR cycle produces mechanical sub-cooling in the subcooled VCR cycle through subcooler using considered refrigerants R134a, R1234yf and R1234ze. The various performance parameters viz. compressor work, COP, exergetic efficiency, total exergy destruction rate and exergy destruction ratio have been explored for different variables viz. degree of sub-cooling, evaporator temperature, effectiveness of subcooler, isentropic efficiency of compressor and condenser temperature.

A set of computer code has been formulated in Engineering Equation Solver (EES) (Klein and Alvarado [22]) to perform the analysis. The steady state governing equations, based on first and second law of thermodynamics have been developed.

Energy Analysis

The first law based work, energy and mass conservation principles are presented in equations (1)-(2).

$$\sum \dot{Q} - \sum \dot{W} = \sum \dot{m}_o h_o - \sum \dot{m}_i h_i \tag{1}$$

$$\sum \dot{m}_i - \sum \dot{m}_o = 0 \tag{2}$$

where $\dot{Q}$, $\dot{W}$ and $\dot{m}$ are the rate of heat, work and mass transfer crossing the boundary of the system respectively.

The coefficient of performance (COP) of simple, dedicated mechanically subcooled, subcooler and overall VCR cycles are given by the equation (3)-(9).

$$COP_{SVCR} = \frac{\dot{Q}_e}{\dot{W}_{comp}} \tag{3}$$

where $\dot{Q}_e$ is the net refrigeration effect produced and $\dot{W}_{comp}$ is the actual work of the compressor.

$$COP_{DSC} = \frac{\dot{Q}_{e,DSC}}{\dot{W}_{comp1}} \tag{4}$$

$$\dot{Q}_{e,DSC} = \dot{m}_{r,DSC}(h_1 - h_4) \tag{5}$$

$$\dot{W}_{comp1} = \dot{m}_{r,DSC}(h_2 - h_1) \tag{6}$$

$$COP_{SC} = \frac{\dot{Q}_{e,SC}}{\dot{W}_{comp2}} \tag{7}$$

$$\dot{W}_{comp2} = \dot{m}_{r,SC}(h_6 - h_5) \tag{8}$$

$$COP_{OC} = \frac{\dot{Q}_{e,DSC}}{\dot{W}_{comp1}+\dot{W}_{comp2}} \tag{9}$$

where $\dot{Q}_{e,DSC}$, $\dot{Q}_{e,SC}$, $\dot{W}_{comp1}$, $\dot{W}_{comp2}$, $\dot{m}_{r,DSC}$ and $\dot{m}_{r,SC}$ are net refrigerant effect, net compressor work and mass flow rate of dedicated mechanically subcooled and subcooler VCR cycles respectively.

The effectiveness of subcooler ε_{SC} is the ratio of minimum heat transfer to the maximum heat transfer.

$$\varepsilon_{SC} = \frac{T_5 - T_8}{T_3 - T_8} \tag{10}$$

where ε_{SC} is the effectiveness of subcooler and T_3,T_5 and T_8 are temperatures at respective state points.

The energy balance in the subcooler is given by equation (11)

$$\dot{m}_{r,DSC}\,(h_3 - h_{3a}) = \dot{m}_{r,SC}\,(h_5 - h_8) \tag{11}$$

Exergy Analysis

The second law of thermodynamics expresses the concept of exergy. The exergy is defined as the degree measure of quality or usefulness of energy to impact the environment. The maximum useful work which can be extracted from a system as it reversibly comes into equilibrium with its environment. (Bejan et al. (1996), Arora and Kaushik [12]).

For the refrigerant flowing in a refrigerating system, the exergy is defined as follows (Bejan et al. [20]):

$$\dot{E} = \dot{m}_r[(h - h_0) - T_0(s - s_0)] \tag{12}$$

where h_0 and s_0 are the enthalpy and entropy values of the refrigerant at dead state pressure P_0 and temperature T_0.

Exergy destruction($\dot{E}_D$) or internal exergy destruction losses which are caused by irreversibilites of the system is the algebric sum of total exergy at the inlet and outlet of the system. General exergy balance is given by (Dincer and Kanoglu [21]):

$$\dot{E}_D = \dot{E}_{in} - \dot{E}_{out} \tag{13}$$

where $\dot{E}_D$ is the rate of exergy destruction and $\dot{E}_{in}$ and $\dot{E}_{out}$ are the total exergy transferred by heat, work and mass.

Exergy Destruction ($\dot{E}_D$) in the Components of Dedicated Mechanically Subcooled Vapour Compression Refrigeration Cycle (DSC)

The exergy destruction in each component of the dedicated mechanically subcooled vapour compression refrigeration cycle is given by the following equations:

Evaporator 1

$$\left(\dot{E}_D\right)_{e,DSC} = \dot{E}_{X_4} + \dot{Q}_{e,DSC}\left(1 - \frac{T_0}{T_{b,DSC}}\right) - \dot{E}_{X_1} = \dot{m}_{r,DSC}(h_4 - T_0 s_4) + \dot{Q}_{e,DSC}\left(1 - \frac{T_0}{T_b}\right) - \dot{m}_{r,DSC}(h_1 - T_0 s_1) \tag{14}$$

Compressor 1

$$\left(\dot{E}_D\right)_{comp1} = \dot{E}_{X_1} + \dot{W}_{comp1} - \dot{E}_{X_2} = \dot{m}_{r,DSC}(\mathrm{h}_1 - T_0 s_1) + \dot{W}_{comp1} - \dot{m}_{r,DSC}(\mathrm{h}_2 - T_0 s_2) \quad (15)$$

Condenser 1

$$\left(\dot{E}_D\right)_{cond1} = \dot{E}_{X_2} - \dot{E}_{X_3} = \dot{m}_{r,DSC}(h_2 - T_0 s_2) - \dot{m}_{r,DSC}(h_3 - T_0 s_3) \quad (16)$$

Expansion Valve 1

$$\left(\dot{E}_D\right)_{Ex.\ Valve1} = \dot{E}_{X_3} - \dot{E}_{X_{3a}} = \dot{m}_{r,DSC}(h_{3a} - T_0 s_{3a}) - \dot{m}_{r,DSC}(h_4 - T_0 s_4) = \dot{m}_{r,DSC} T_0 (s_4 - s_{3a}) \quad (17)$$

Subcooler

$$\left(\dot{E}_D\right)_{SC} = (\dot{E}_{X_8} - \dot{E}_{X_5}) + (\dot{E}_{X_3} - \dot{E}_{X_{3a}}) = \dot{m}_{r,SC}\{(h_8 - h_5) - T_0(s_8 - s_5)\} + \dot{m}_{r,DSC}\{(h_3 - h_{3a}) - T_0(s_3 - s_{3a})\} \quad (18)$$

Compressor 2

$$\left(\dot{E}_D\right)_{comp2} = \dot{E}_{X_5} + \dot{W}_{comp2} - \dot{E}_{X_6} = \dot{m}_{r,SC}(h_5 - T_0 s_5) + \dot{W}_{comp2} - \dot{m}_{r,SC}(h_6 - T_0 s_6) \quad (19)$$

Condenser 2

$$\left(\dot{E}_D\right)_{cond2} = \dot{E}_{X_6} - \dot{E}_{X_7} = \dot{m}_{r,SC}(h_6 - T_0 s_6) - \dot{m}_{r,SC}(h_7 - T_0 s_7) \quad (20)$$

Expansion Valve 2

$$\left(\dot{E}_D\right)_{Ex.\ valve2} = \dot{E}_{X_7} - \dot{E}_{X_8} = \dot{m}_{r,SC}(h_7 - T_0 s_7) - \dot{m}_{r,SC}(h_8 - T_0 s_8) = \dot{m}_{r,SC} T_0 (s_8 - s_7) \quad (21)$$

Evaporator 2

$$\left(\dot{E}_D\right)_{e,Sc} = \dot{E}_{X_8} + \dot{Q}_{e,SC}\left(1 - \frac{T_0}{T_{b,SC}}\right) - \dot{E}_{X_5} = \dot{m}_{r,SC}(h_8 - T_0 s_8) + \dot{Q}_{e,SC}\left(1 - \frac{T_0}{T_{b,SC}}\right) - \dot{m}_{r,SC}(h_5 - T_0 s_5) \quad (22)$$

where $\dot{E}_D$, $\dot{E}_X$, $\dot{m}_r$ and T_b are the exergy destruction rate, exergy rate, mass flow rate and temperature of space to be cooled for the dedicated mechanically subcooled and subcooler cycle respectively. T_0, h_0 and s_0 are the temperature, enthalpy and entropy of dead state.

Total Exergy Destruction

Total exergy destruction of the system is the sum of exergy destruction in each component of the system. The total exergy destruction in dedicated mechanically subcooled vapour compression refrigeration cycle (DSC) is given by:

$$\sum(\dot{E}_D)_{DSC} = (\dot{E}_D)_{e,DSc} + (\dot{E}_D)_{comp1} + (\dot{E}_D)_{cond1} + (\dot{E}_D)_{Ex.\ valve1} \tag{23}$$

$$\sum(\dot{E}_D)_{SC} = (\dot{E}_D)_{e,Sc} + (\dot{E}_D)_{comp2} + (\dot{E}_D)_{cond2} + (\dot{E}_D)_{Ex.\ valve2} \tag{24}$$

$$\sum(\dot{E}_D)_{OC} = (\dot{E}_D)_{e,DSc} + (\dot{E}_D)_{comp1} + (\dot{E}_D)_{cond1} + (\dot{E}_D)_{Ex.\ valve1} + (\dot{E}_D)_{SC} + (\dot{E}_D)_{comp2} + (\dot{E}_D)_{cond2} + (\dot{E}_D)_{Ex.\ valve2} \tag{25}$$

Exergetic Efficiency

The exergetic efficiency is the ratio of total exergy recovered to the total exery supplied and is given by (Dincer and Kanoglu (2010)):

$$\eta_{ex} = \frac{\sum \dot{E}_{out}}{\sum \dot{E}_{in}} = 1 - \frac{\sum \dot{E}_D}{\sum \dot{E}_{in}} \tag{26}$$

where η_{ex} is the exergetic efficiency of the cycle. $\sum \dot{E}_{out}$, $\sum \dot{E}_{in}$ and $\sum \dot{E}_D$ are the total exergy recovered, supplied and destructed respectively.

The exergetic efficiency of vapour compression refrigeration system is defined as the ratio of the exergy of heat absorbed in the evaporator from the space to be cooled at temperature T_b to the actual compressor work input ($\dot{W}_{comp}$) (Arora and Kaushik [12]). Exergy efficiency can also be defined as The exergetic efficiency is the ratio of exery in products to the exergy of fuel. The exergetic efficiency of dedicated mechanically subcooled, subcooler and overall VCR cycles are given by the equations (27)-(29).

$$\eta_{ex\ (DSC)} = \frac{\left|\dot{Q}_{e,DSC}\left(1-\frac{T_0}{T_{b,DSC}}\right)\right|}{\dot{W}_{comp1}} \tag{27}$$

$$\eta_{ex\ (SC)} = \frac{\left|\dot{Q}_{e,SC}\left(1-\frac{T_0}{T_{b,SC}}\right)\right|}{\dot{W}_{comp2}} \tag{28}$$

$$\eta_{ex\ (OC)} = \frac{\left|\dot{Q}_{e,OC}\left(1-\frac{T_0}{T_{b,DSC}}\right)\right|}{\dot{W}_{comp1}+\dot{W}_{comp2}} \tag{29}$$

where T_0 is the ambient or dead state temperature.

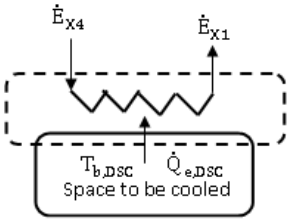


Figure 3. Exergy balance in evaporator of DSC and the space to be cooled [23]

Exergy Destruction Ratio (EDR)

The EDR is the total exergy destruction in the system to the exergy in the products (Arora and Kaushik [12]) and is given by equation (30).

$$EDR = \frac{\dot{E}_{D,total}}{\dot{E}_P} = \frac{1}{\eta_{ex}} - 1 \tag{30}$$

The exergy destruction ratio for the dedicated mechanically subcooled, subcooler and overall VCR cycles are given in following (31)-(33).

$$EDR_{DSC} = \frac{\sum(\dot{E}_D)_{DSC}}{\dot{E}_{P,DSC}} = \frac{\sum(\dot{E}_D)_{DSC}}{\left|\dot{Q}_{e,DSC}\left(1-\frac{T_0}{T_{b,DSC}}\right)\right|} \tag{31}$$

$$EDR_{SC} = \frac{\sum(\dot{E}_D)_{SC}}{\dot{E}_{P,SC}} = \frac{\sum(\dot{E}_D)_{SC}}{\left|\dot{Q}_{e,SC}\left(1-\frac{T_0}{T_{b,SC}}\right)\right|} \tag{32}$$

$$EDR_{OC} = \frac{\sum(\dot{E}_D)_{OC}}{\dot{E}_{P,DSC}} = \frac{\sum(\dot{E}_D)_{OC}}{\left|\dot{Q}_{e,DSC}\left(1-\frac{T_0}{T_{b,DSC}}\right)\right|} \tag{33}$$

Mathematical Modelling

The present work investigates the performance of dedicated mechanically subcooled vapour compression refrigeration cycle (DSC) on the bassis of first and second law of thermodynamics. The mathematical fomulation of thermodynamic relations have been carried out considering mass, energy and work conservation principles for system and system components.

Table 1. Input variable considered except where the variation of these system variables involved (Arora and Kaushik [12])

S.No.	Input variables	Values
1.	Evaporator1 temperature ($T_{e,\,DSC}$)	-10^0C
2.	Degree of subcooling ($(\Delta T)_{sc}$)(3-3a)	5^0C
3.	Condenser1,2 temperature ($T_{c,DSC}$ and $T_{c,SC}$), assuming summer conditions in a country near the equator line.	50^0C
4.	Isentropic efficiency of compressor ($\eta_{comp1}, \eta_{comp2}$)	80%
5.	Effectiveness of subcooler (ε_{SC})	0.8
6.	Net refrigerating effect ($\dot{Q}_{e,DSC}$)	3.5167 kW
7.	Ambient or dead state temperature (T_0) and pressure (P_0)	25^0C and 101.325 kPa

The following assumptions have been considered to formulate the complex thermodynamic model of dedicated mechanically subcooled vapour compression refrigeration cycle (DSC) except where the variation of physical parameters involved.

- It is assumed that the state of refrigerant is dry and saturated at the entry of compressor at state points 1 and 5.
- Difference between evaporator temperature and the temperature of space to be cooled ($T_{b,\,DSC}$ -$T_{e,\,DSC}$, $T_{b,SC}$ -$T_{e,SC}$) is 5 ^{0}C.
- The range of variation of evaportor1 temperature ($T_{e,\,DSC}$) is -20 ^{0}C to 10 ^{0}C.
- The range of variation of condenser1 temperature ($T_{e,\,SC}$) is 30 ^{0}C to 50 ^{0}C.

- The heat losses and pressure losses from the system and system components are negligible. The whole system operates in steady state condition.
- The range of degree of subcooling $(\Delta T)_{SC}$ is 5 to 30 ^{0}C.
- The pressure drop in evaporator (δ_e) and condenser (δ_c) is assumed to be zero.

RESULTS AND DISCUSSION

Energy and exergy analysis of dedicated mechanically subcooled vapour compression refrigeration cycle (DSC) have been performed using a computer software based program in Engineering Equation Solver (EES) (Klein and Alvarado [22]) which has been shown in Figure 1(a) and 2(b). The considered refrigerants for the analysis are R134a, R1234yf and R1234ze. The computation of the various performance parameters viz. compressor work, COP, exergetic efficiency, total exergy destruction rate and exergy destruction ratio (EDR) have been done by calling builtin functions (i.e. specific entropy, specific enthalpy, temperature and pressure etc.) of the EES library.

The results from the present analysis have been compared with the results of Arora et al. [12]. It is observed that the difference in the value of COP and exergetic efficiency is within a range of ±1%. The difference in results may be due to the pressure drop in evaporator and condenser is assumed to be zero.

Effect of Subcooling

The effect of degree of sub-cooling $(\Delta T)_{sc}$ on various performance parameters viz. compressor work $(\dot{W}_C)$, COP, exergetic efficiency (η_{ex}), total exergy destruction rate $(\dot{E}_{D,\ Total})$ and exergy destruction ratio (EDR) has been illustrated in Figure 4(a) to 4(d).

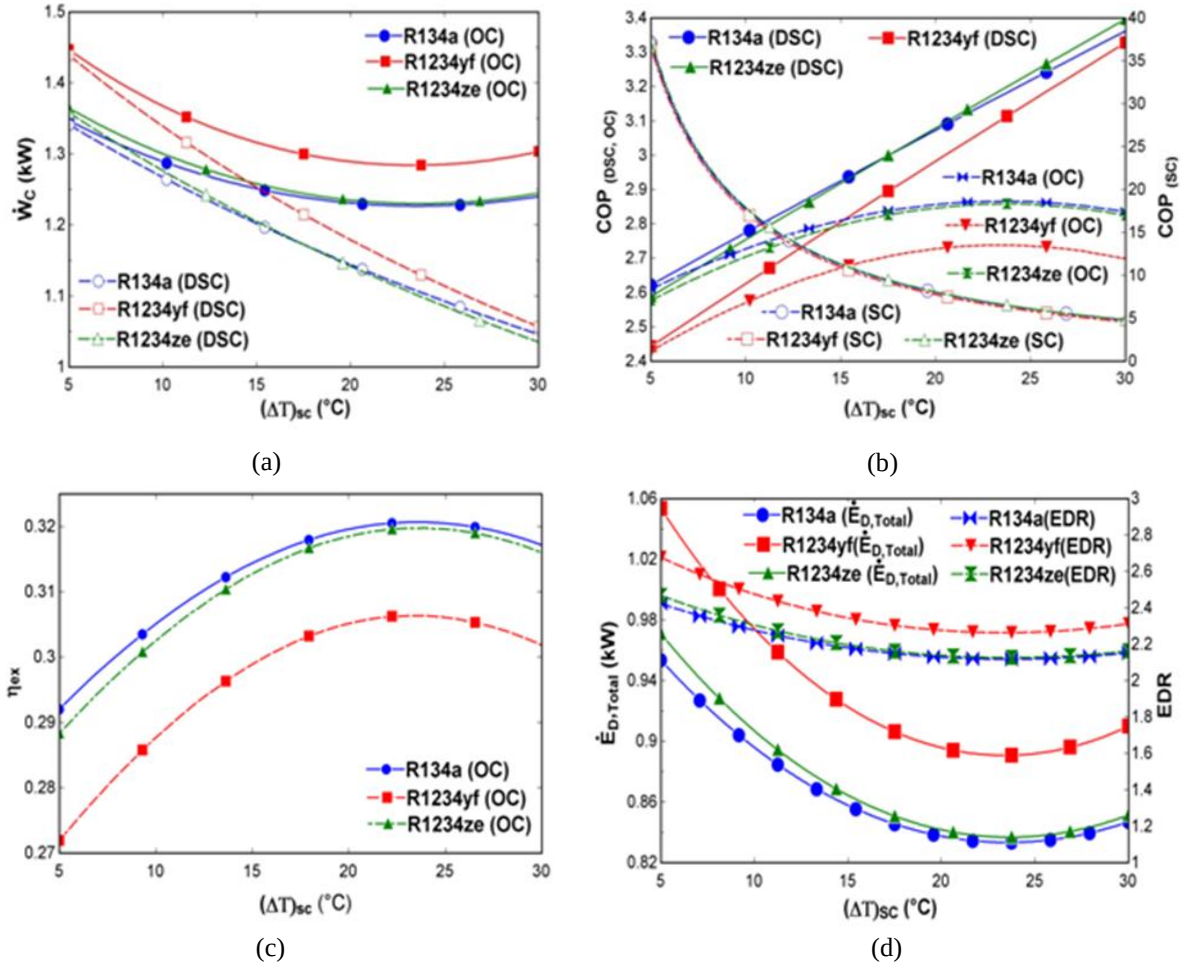


Figure 4. (a) Variation in compressor work $(\dot{W}_C)$ with degree of subcooling (ΔT_{sc}), (b) effect of degree of subcooling (ΔT_{sc}) on COP, (c) Effect of degree of subcooling (ΔT_{sc}) on exergetic efficiency (ηex), (d) variation in total exergy destruction rate (E ̇_(D,Total)) and exergy destruction ratio (EDR) with subcooling (ΔT_{sc}) of OC

Figure 4(a) represents the effect of sub-cooling $(\Delta T)_{sc}$ on compressor work $(\dot{W}_C)$. For 1 Ton refrigeration, compressor work of dedicated mechanically subcooled (DSC) $(\dot{W}_{Comp1})$ decreases with increase in degree of subcooling (DOS) because the mass flow rate of refrigerant decreases with increase in degree of subcooling for constant refrigeration load of 3.5167kW. However, the compressor work first decreases and then increases for overall cycle (OC) with increase in degree of sub-cooling for the considered refrigerants. This is due to that the compressor work of OC $(\dot{W}_{Comp1} + \dot{W}_{Comp2})$ decreases for low values of DOS as the value of $\dot{W}_{Comp1}$ decreases and the values of $\dot{W}_{Comp2}$ are very small. For the higher values of DOS, the compressor work of subcooler cycle $(\dot{W}_{Comp2})$ increases while $\dot{W}_{Comp1}$ decreases. However the compressor work for the simple VCR cycle is is higher than that of the DSC. The optimize value of degree of sub-cooling lies within 20-25^{0}C for overall cycle. It has been observed that the compressor work is minimum for R1234ze and maximum for R1234yf for DSC and OC.

Figure 4(b) shows the variation of COP with degree of sub-cooling for the considered refrigerants R134a, R1234yf and R1234ze. The COP of the dedicated mechanically subcooled cycle increases and the COP of subcooler cycle decreases with increase in DOS for the considered refrigerants. However, the COP of the overall cycle first increases, maximise for optimum value of DOS which lies within 20-25^{0}C and then decreases with increase in DOS. In order to increase DOS, The temperature and pressure of evaporator of subcooler cycle has to achieved lower values which reduces the COP of the subcooler cycle while for the constant refrigerating load of $\dot{Q}_{e,DSC}$ = 3.5167kW, the compressor work $(\dot{W}_{Comp1})$ of DSC decreases linearly and the COP of DSC is $\frac{\dot{Q}_{e,DSC}}{\dot{W}_{comp1}}$ ("Eq.(4)"). Thus the COP of the DSC increases linearly with increase in DOS. The COP of OC is $\frac{\dot{Q}_{e,DSC}}{\dot{W}_{comp1}+\dot{W}_{comp2}}$ ("Eq.(9)") which increases with increase in COP of DSC and decreases with decrease in COP of subcooler cycle due to increase in compressor work of SC with increase in DOS. Thus the COP of OC first increases (as COP of DSC increases faster) reaches to maximum value and then decreases (as COP of SC cycle decreases faster) for the refrigerant considered. It is observed that the COP is maximum for R1234ze and minimum for R1234yf. The COP of simple VCR cycle is lower than the DSC and OC.

Figure 4(c) illustrates the effect of degree of sub-cooling (ΔT_{sc}) on exergetic efficiency $(\eta_{ex\ (OC)})$ of overall cycle. The exergetic efficiency of the overall cycle first increases reaches to a maximum value and then decreases with increase in DOS for the considered refrigerants. The optimum value of degree of sub-cooling lies between 20 to 25^{0}C for which the value of exergetic efficiency is maximum. The exergetic efficiency of overall cycle is given by $\eta_{ex\ (OC)} = \frac{\left|\dot{Q}_{e,OC}\left(1-\frac{T_0}{T_{b,DSC}}\right)\right|}{\dot{W}_{comp1}+\dot{W}_{comp2}}$ (eq. (29)). In which the term $\frac{|\dot{Q}_{e,OC}|}{\dot{W}_{comp1}+\dot{W}_{comp2}}$ is the $COP_{(OC)}$ which first increases, reaches to a maximum value and then decreases with increase in DOS (Figure 4(b)) and the term $\left|\left(1-\frac{T_0}{T_{b,DSC}}\right)\right|$ remains constant with increase in DOS because the value of T_0 and $T_{b,DSC}$ remains constant with increase in DOS. Hence the value of $\eta_{ex\ (OC)}$ first increases, reaches to a maximum value and then decreases. The optimum value of DOS lies between 20 to 25^{0}C for the considered refrigerants. Hence the exergetic efficiency of the OC follows the trend of COP. It can be observed that $\eta_{ex\ (OC)}$ is maximum for R134a and minimum for R1234yf.

Figure 4(d) shows the variation in total exergy destruction rate $(\dot{E}_{D,\ Total})$ and exergy destruction ratio (EDR) with degree of subcooling (ΔT_{sc}) for overall cycle. The value of $\dot{E}_{D,\ Total}$ and EDR first decreases, reaches to a minimum value and then increases with increase in DOS for the refrigerants considered. The value of $\dot{Q}_{e,OC}$ first increases reaches to a maximum value and then decreases with increase in DOS (Figure 4(b)). Therefore, the exergy recovered from the system and system components first increases, reaches to a maximum value and then decreases with increase in DOS. Consequently, the $\dot{E}_{D,\ Total}$ first decreases, reaches to a minimum value and then increases. EDR is given by $\frac{\Sigma(\dot{E}_D)_{OC}}{\left|\dot{Q}_{e,DSC}\left(1-\frac{T_0}{T_{b,DSC}}\right)\right|}$ (eq. (33)) which is proportional to $\dot{E}_{D,\ Total}$ and reciprocal to $\left|\dot{Q}_{e,DSC}\left(1-\frac{T_0}{T_{b,DSC}}\right)\right|$. Hence EDR first decreases, reaches to a minimum value and

then increases with increase in DOS (ΔT_{sc}). It can also be observed that the values of $\dot{E}_{D,\ Total}$ and EDR are minimum for R134a and maximum for R1234yf.

Effect of Evaporator Temperature

The effect of evaporator1 temperature ($T_{e,\ DSC}$) on various performance parameters viz. compressor work $(\dot{W}_C)$, degree of subcooling $(\Delta T)_{sc}$, COP, exergetic efficiency (η_{ex}), total exergy destruction rate $(\dot{E}_{D,\ Total})$ and exergy destruction ratio (EDR) has been shown in Figure 5(a) to 5(d).

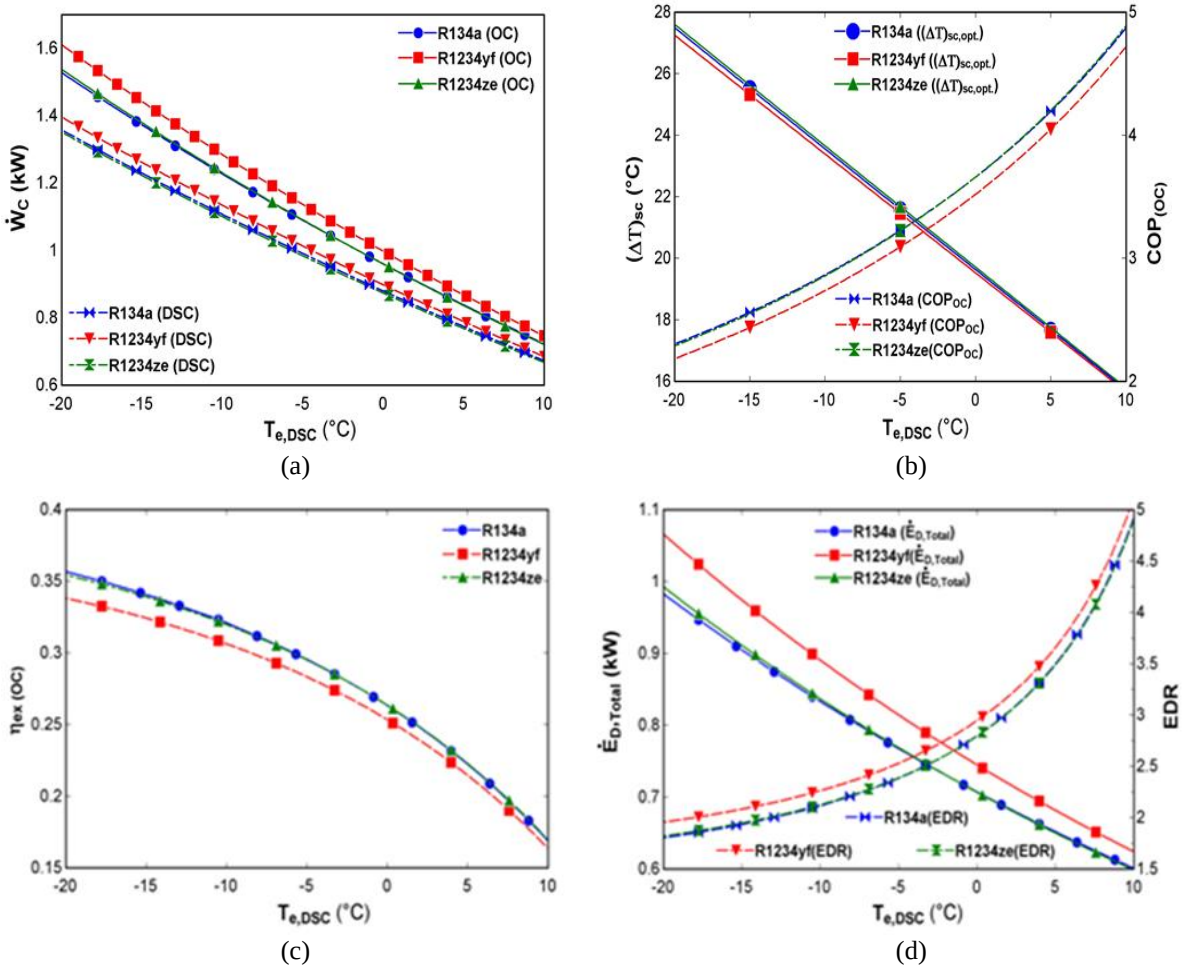


Figure 5. (a) Variation in compressor work$(\dot{W}_C)$ with evaporator1 Temperature ($T_{e,\ DSC}$), (b) effect of evaporator1 temperature ($T_{e,DSC}$) on degree of subcooling (ΔT_{sc}) and COP, (c) effect of evaporator1 temperature ($T_{e,DCS}$) on exergetic efficiency (η_{ex}) of overall cycle, (d) variation in total exergy destruction rate $(\dot{E}_{D,Total})$ and exergy destruction ratio (EDR) with evaporator1 temperature ($T_{e,DCS}$) of overall cycle

Figure 5(a) shows the varition in compressor work $(\dot{W}_C)$ with evaporator1 temperature ($T_{e,\ DSC}$) for dedicated mechanically subcooled and overall VCR cycle. The compressor work of DSC $(\dot{W}_{Comp1})$ and OC $(\dot{W}_{Comp1} + \dot{W}_{Comp2})$ decreases with increase in evaporator1 temperature ($T_{e,\ DSC}$) for considered refrigerants. As the evaporator1 temperature increases, the pressure ratio between evaporator and condenser decreases and for constant refrigeration load of 3.5167kW, the mass flow rate of refrigerant also decreases. Moreover, the evaporator2 temperature of sub-cooler cycle also increases. Hence the compressor work of DSC and OC decreases with increase in evaporator1 temperature. It is observed that the value of compressor work is minimum for R1234ze and maximum for R1234yf.

Figure 5(b) illustrates the effect of evaporator1 temperature ($T_{e,\ DSC}$) on DOS $(\Delta T)_{sc}$ and COP for OC. The value of DOS decreases and COP increases with increase in evaporator1 temperature for the refrigerant considered. For the constant refrigeration load of 3.5167kW, lower amount of sub-cooling is required in the sub-

cooler to the refrigerant of DSC. Thus the DOS decreases with increase in $T_{e,\,DSC}$. However, the compressor work decreases as shown in Figure 5(a), the value of COP increases. It has been observed that the value of DOS and COP are maximum for R1234ze and minimum for R1234yf.

Figure 5(c) depicts the effect of evaporator1 temperature ($T_{e,\,DSC}$) on exergetic efficiency (η_{ex}) of OC for considered refrigerants. The exergetic efficiency is given by $COP_{(OC)} \times \left|\left(1-\frac{T_0}{T_{b,DSC}}\right)\right|$ in which the value of $COP_{(OC)}$ increases (Figure 5(b)) and the value of term $\left|\left(1-\frac{T_0}{T_{b,DSC}}\right)\right|$ decreases with increase in evaporator1 temperatue as ($T_{b,\,DSC}$ -$T_{e,\,DSC}$ =5^0C). `Hence, the value of $\eta_{ex\,(OC)}$ decreases with increase in evaporator1 temperature. It can be observed that the value of $\eta_{ex\,(OC)}$ is maximum for R134a and minimum for R1234yf.

Figure 5(d) illustrates the variation in total exergy destruction rate $\left(\dot{E}_{D,\ Total}\right)$ and exergy destruction ratio (EDR) with evaporator1 temperature ($T_{e,\,DSC}$) of OC. The decreases with increase in $T_{e,\,DSC}$. It is observed that the value of $\dot{E}_{D,\ Total}$ and EDR are minimum for R1234ze and maximum for R1234yf. $\dot{E}_{D,\ Total}$ increases and EDR decreases with increase in $T_{e,\,DSC}$ for refrigerant considered. As the exergetic efficiency decreases with increase in evaporator1 temperature (Figure 5(c)), the rate of exergy recovered of system component decreases. By which the $\dot{E}_{D,\ Total}$ decreases with increase in $T_{e,\,DSC}$. However, the exergy of product decreases with increase in $T_{e,\,DSC}$ because the term $\left|\left(1-\frac{T_0}{T_{b,DSC}}\right)\right|$ decreases due to which the EDR of the overall cycle.

Effect of Condenser Temperature

The effect of degree of condenser1,2 temperature ($T_{c,\,DSC} = T_{c,\,SC}$) on various performance parameters viz. compressor work $\left(\dot{W}_C\right)$, COP, exergetic efficiency (η_{ex}), total exergy destruction rate $\left(\dot{E}_{D,\ Total}\right)$ and exergy destruction ratio (EDR) has been depicted in Figure 6(a) to 6(d).

Figure 6(a) illustrates the effect of condenser1,2 temperature ($T_{c,\,DSC} = T_{c,\,SC}$) on compressor work ($\dot{W}_{C,SVCR}, \dot{W}_{Comp1}, \dot{W}_{Comp2}$ and $(\dot{W}_{Comp1} + \dot{W}_{Comp2})$) of simple vapour compression refrigeration cycle (SVCR), DSC, SC and OC for refrigerants considered. The values of compressor work increases with increase in condenser temperature. As the condenser temperature increases, the pressure ratio between evaporator and condenser increases for which more work has been done by the compressor. Therefore, the value of $\dot{W}_C$ increases with increase in T_c. It is also observed that the value of $\dot{W}_C$ is minimum for R1234ze and maximum for R1234yf in DSC, OC and SVCR.

Figure 6(b) shows the effect of condenser1,2 temperature ($T_{c,\,DSC} = T_{c,\,SC}$) on COP of DSC, OC, SC and SVCR for considered refrigerants. The COP of DSC, OC and SVCR decreases with increase in $T_{c,\,DSC} = T_{c,\,SC}$. This is due to that for constant refrigeration load of 3.5167kW, the compressor work increases with increase in condenser temperature. Thus the COP of DSC, OC and SVCR decreases with increase in condenser temperature. However, the COP of sub-cooler VCR cycle (SC) increases with increase in condenser temperature. The rate of heat of heat transfer to evaporator2 increases in the sub-cooler, consequently the refrigeration load of sub-cooler cycle increases and hence COP of sub-cooler cycle increases with increase in condenser temperature. It has also been observed that the COP is maximum for R1234ze and minimum for R1234yf.

Figure 6(c) presents the variation in exergetic efficiency (η_{ex}) with condenser temperature ($T_{c,\,DSC} = T_{c,\,SC}$) of DSC, OC and SVCR cycles for considered refrigerants. The value of η_{ex} decreases with increase in $T_{c,\,DSC} = T_{c,\,SC}$. Exergetic efficiency is given by the relation, $\eta_{ex} = COP \times \left|\left(1-\frac{T_0}{T_b}\right)\right|$ in which the term $\left|\left(1-\frac{T_0}{T_b}\right)\right|$ remains constant and the value of COP decreases with increase in condenser temperature. Hence the value of η_{ex} decreases with increase in $T_{c,\,DSC} = T_{c,\,SC}$. It can also be observed that the value of η_{ex} is maximum for R1234ze and minimum for R1234yf.

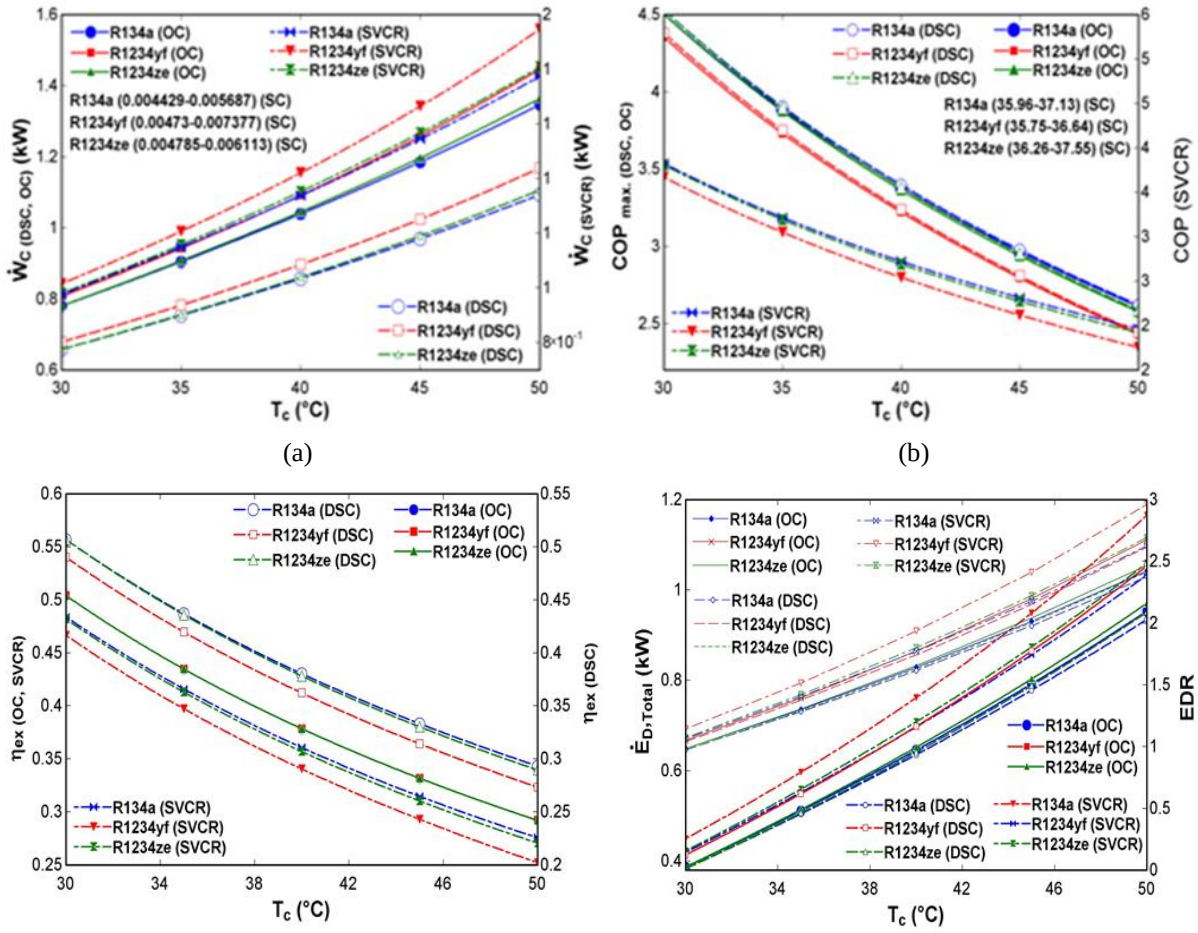


Figure 6. (a) Variation in compressor work $(\dot{W}_C)$ with condenser temperature (T_c), (b) effect of condenser temperature ($T_{e,DSC}$) on COP, (c) Effect of condenser temperature (T_c) on exergetic efficiency (η_{ex}), (d) Variation in total exergy destruction rate $(\dot{E}_{D,Total})$ and exergy destruction ratio (EDR) with condenser temperature (T_c)

Figure 6(d) depicts the effect of condenser temperature ($T_{c,\,DSC} = T_{c,\,SC}$) on total exergy destruction rate $(\dot{E}_{D,Total})$ and exergy destruction ratio (EDR) of DSC, OC and SVCR cycles. The value of $\dot{E}_{D,\ Total}$ and EDR increases with increase in $T_{c,\,DSC} = T_{c,\,SC}$. As the condenser temperature increases, the exergy destruction rate in system components increases as well as the exergy of products $\left(\left|\dot{Q}_{e,DSC}\left(1-\frac{T_0}{T_{b,DSC}}\right)\right|\right)$ decreases because the value of COP decreases (Figure 6(b)). Hence the total exergy destruction rate and EDR increase with increase in $T_{c,\,DSC} = T_{c,\,SC}$. It has been seen that the value of $\dot{E}_{D,\ Total}$ and EDR is minimum for R134a and maximum for R1234yf.

CONCLUSIONS

The current work presents the energy and exergy analysis of dedicated mechanically sub-cooled vapour compression refrigeration cycle (DSC). The comparison of DSC, overall cycle (OC), sub-cooler cycle (SC) and simple vapour compression refrigeration cycle (SVCR) has also been considered for R134a, R1234ya and R1234ze using a computer software based program in EES. The theoretical performance analysis involves the variation of DOS, evaporator temperature and condenser temperature to check the performance parameters viz. compressor work, COP, exergetic efficiency, and total exergy destruction rate and exergy destruction ratio. The main concluding points of the current work are as follows:

- The COP of dedicated mechanically sub-cooled VCR cycle (DSC) and overall cycle (OC) is higher than the simple VCR cycle.

- The optimum value of DOS lies between 20^0C to 25^0C for which $\dot{W}_C$, COP and η_{ex} are maximum in OC.
- The COP of OC improves while η_{ex} goes down with increase in evaporator temperature.
- The high condenser temperature reduces the performance (COP andη_{ex}) of the cycle.
- Higher condenser temperature promotes exergy destruction while exergy destruction reduces for lower values of evaporator temperature. The optimum value of DOS lies between 20^0C to 25^0C.
- The performance of R1234ze is better than R1234yf and comparable to R134a.
- Finally, it is inferred that the dedicated mechanically sub-cooled VCR cycle performs better than that of simple VCR cycle and the performance of HFO R1234ze is better than R1234yf.

NOMENCLATURE

COMP	Compressor
Cond	Condenser
COP	Coefficient of performance
DOS	Degree of sub-cooling
DSC	Dedicated mechanically sub-cooled VCR cycle
$(\Delta T)_{sc}$	Degree of subcooling
$\dot{E}_D$	Exergy destruction rate (kW)
ED	Exergy destruction
EDR	Exergy destruction ratio
EES	Energy equation solver
$\dot{E}$	Rate of exergy (kW)
E.V.	Expansion valve
GWP	Global warming potential
h	Specific enthalpy (kJ/kg)
HFC	Hydrofluorocarbon
HFO	Hydrofluoroolefin
$\dot{m}_r$	Mass flow rate of refrigerant (kg/s)
ODP	Ozone depleting potential
P	Pressure (kappa)
SVCR	Simple vapour compression refrigeration
$\dot{Q}_e$	Rate of net refrigerating effect (kW)
s	Specific entropy (kJ/kg^{0}C)
T	Temperature (^{0}C)
T_b	Boundary temperature (^{0}C)
T_e	Evaporator temperature (^{0}C)
T_0	Dead state temperature (^{0}C)
VCR	Vapour compression refrigeration
$\dot{W}$	Work rate (kW)
ε	Effectiveness
η	Efficiency
Σ	represents summation
0	Dead state
C	Condenser
Comp	Compressor
Cond	Condenser
e	Evaporator
ex	Exergetic
SC	Su-cooling cycle
OC	Overall cycle
i	Input
o	Output

REFERENCES

[1] Parliament, E., and CotE, U. (2014). Regulation (EU) No 517/2014 of The European Parliament and of the council of 16 April 2014 on fluorinated greenhouse gases and repealing Regulation (EC) No 842/2006. Official J. Eur. Union, 57, 195-230.

[2] Zubair, S. M. (1990). Improvement of refrigeration/air-conditioning performance with mechanical sub-cooling. Energy, 15(5), 427-433.
[3] Zubair, S. M. (1994). Thermodynamics of a vapor-compression refrigeration cycle with mechanical subcooling. Energy, 19(6), 707-715.
[4] Zubair, S. M., Yaqub, M., and Khan, S. H. (1996). Second-law-based thermodynamic analysis of two-stage and mechanical-subcooling refrigeration cycles. International Journal of Refrigeration, 19(8), 506-516.
[5] Couvillion, R. J. (1988). Analysis of a Vapour-compression Refrigeration System with mechanical subcooling. ASHRAE Trans., 94(2), 641-660.
[6] Qureshi, B. A., Inam, M., Antar, M. A., and Zubair, S. M. (2013). Experimental energetic analysis of a vapor compression refrigeration system with dedicated mechanical sub-cooling. Applied Energy, 102, 1035-1041.
[7] Zubair, S. M. (2001). Thermodynamic optimization of finite time vapor compression refrigeration systems. Energy conversion and management, 42(12), 1457-1475.
[8] Thornton, J. W., Klein, S. A., and Mitchell, J. W. (1994). Dedicated mechanical subcooling design strategies for supermarket applications. International Journal of Refrigeration, 17(8), 508-515.
[9] Arora, A., Arora, B. B., Pathak, B. D., and Sachdev, H. L. (2007). Exergy analysis of a vapour compression refrigeration system with R-22, R-407C and R-410A. International journal of Exergy, 4(4), 441.
[10] Arora, A., Dixit, M., and Kaushik, S. C. (2016). Computation of optimum parameters of a half effect water lithium bromide vapour absorption refrigeration system. Journal of Thermal Engineering, 2(2), 683-692.
[11] Arora, A., Dixit, M., and Kaushik, S. C. (2016). Energy and exergy analysis of a double effect parallel flow LiBr/H2O absorption refrigeration system. Journal of Thermal Engineering, 2(1), 541-549.
[12] Arora, A., and Kaushik, S. C. (2008). Theoretical analysis of a vapour compression refrigeration system with R502, R404A and R507A. International journal of refrigeration, 31(6), 998-1005.
[13] Arora, A., Arora, B. B., Pathak, B. D., and Sachdev, H. L. (2007). Exergy analysis of a vapour compression refrigeration system with R-22, R-407C and R-410A. International journal of Exergy, 4(4), 441.
[14] Llopis, R., Cabello, R., Sánchez, D., and Torrella, E. (2015). Energy improvements of CO2 transcritical refrigeration cycles using dedicated mechanical subcooling. International Journal of Refrigeration, 55, 129-141.
[15] Llopis, R., Nebot-Andrés, L., Sánchez, D., Catalán-Gil, J., and Cabello, R. (2018). Subcooling methods for CO 2 refrigeration cycles. A Review. International Journal of Refrigeration.
[16] Mota-Babiloni, A., Navarro-Esbrí, J., Barragán-Cervera, Á., Molés, F., Peris, B., and Verdú, G. (2015). Commercial refrigeration–an overview of current status. International journal of refrigeration, 57, 186-196.
[17] Kalla, S. K., Arora, B. B., and Usmani, J. A. (2018). Alternative refrigerants for HCFC 22 - A review. Journal of Thermal Engineering, 4(3), 1998-2017.
[18] Kalla, S. K., Arora, B. B., and Usmani, J. A. (2018). Performance analysis of R22 and its substitutes in air conditioners. Journal of Thermal Engineering, 4(1), 1724-1736.
[19] Yataganbaba, A., Kilicarslan, A., and Kurtbaş, İ. (2015). Exergy analysis of R1234yf and R1234ze as R134a replacements in a two evaporator vapour compression refrigeration system. International journal of refrigeration, 60, 26-37.
[20] Bejan, A., Tsatsaronis, G., Moran, M., and Moran, M. J. (1996). Thermal design and optimization. John Wiley and Sons.
[21] Dincer, I., and Kanoglu, M. (2010). Refrigeration systems and applications (Vol. 2). New York: Wiley.
[22] Klein, S. A., and Alvarado, F. (2012) Engineering Equation Solver. F-Chart Software, Middleton, WI. Version 9, 224-3D.
[23] Kotas, T. J. (1995) The exergy method of thermal plant analysis, Reprint ed., Krieger Publishing Company, Krieger Drive, Malabar, Florida-32950.
[24] Dixit, M., Kaushik, S. C., and Arora, A. (2017) Energy and exergy analysis of absoption-compression cascade refrigeration system. Journal of Thermal Engineering, 3(5), 1466-1477.
[25] Dixit, M., Arora, A., and Kaushik, S. C. (2016) Energy and exergy analysis of a waste heat driven cycle for triple effect refrigeration. Journal of Thermal Engineering, 2(5), 954-961.
[26] Dixit, M., Kaushik, S. C., and Arora, A. (2017) Energy and exergy analysis of absoption-compression cascade refrigeration system. Journal of Thermal Engineering, 3(5), 1466-1477.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Unified Graph-Based Multicue Feature Fusion for Robust Visual Tracking

Gurjit Singh Walia, *Senior Member, IEEE*, Himanshu Ahuja, *Student Member, IEEE*, Ashish Kumar, *Member, IEEE*, Nipun Bansal, *Member, IEEE*, and Kapil Sharma, *Member, IEEE*

***Abstract*—Visual tracking is a complex problem due to unconstrained appearance variations and a dynamic environment. The extraction of complementary information from the object environment via multiple features and adaption to the target's appearance variations are the key problems of this paper. To this end, we propose a robust object tracking framework based on the unified graph fusion (UGF) of multicue to adapt to the object's appearance. The proposed cross-diffusion of sparse and dense features not only suppresses the individual feature deficiencies but also extracts the complementary information from multicue. This iterative process builds robust unified features which are invariant to object deformations, fast motion, and occlusion. Robustness of the unified feature also enables the random forest classifier to precisely distinguish the foreground from the background, adding resilience to background clutter. In addition, we present a novel kernel-based adaptation strategy using outlier detection and a transductive reliability metric. The adaptation strategy updates the appearance model to accommodate variations in scale, illumination, and rotation. Both qualitative and quantitative analyses on benchmark video sequences from OTB-50, OTB-100, VOT2017/18, and UAV123 show that the proposed UGF tracker performs favorably against 18 other state-of-the-art trackers under various object tracking challenges.**

***Index Terms*—Adaptive appearance model, feature fusion, outlier detection, visual tracking.**

I. Introduction

TRACKING objects in video sequences have emerged as one of the most challenging problems in the field of computer vision. Generally, the object tracking framework aims at keeping track of the targeted object in subsequent frames while catering to various tracking challenges. The challenges in object tracking arise due to changes in object appearance, such as its deformation, scale variations, fast motion, and inplane and out-of-plane rotation, and due to a dynamic environment, such as illumination variations, occlusion, and background clutter. Despite this, object tracking has been investigated extensively due to its wide range of applications, *viz.*, military surveillance, robotics, medical imaging, augmented reality, and human–computer interaction.

Recently, multicue-based object tracking solutions have been investigated and reviewed [1]. Typically, multicue-based trackers consider complementary visual cues so that if one cue fades, the other cue can compensate. These trackers can fuse the extracted information at the score or the feature level. The score-level approach uses classifier scores for segmentation of targets from the background. The multiscale spatiotemporal context tracker [2], multicue tracker [3], and hash tracker [4] considered score-level fusion. On the other hand, feature-level fusion combines different feature vectors extracted from individual cues. Lan *et al.* [5] used a sparse representation for feature-level fusion to remove unreliable features dynamically. In [6], color and histogram of orientation gradient (HOG) features are considered for object representation. Generally, trackers based on feature-level fusion fail to capture the nonlinear variations in illumination, deformations, and occlusion of the targeted object due to offcourse updating of the tracker. Despite recent investigations in this direction, adaptive nonlinear feature fusion to build a robust appearance model, invariant to the dynamic environment remains a challenge.

The tracker's appearance model keeps a record of the visual changes in the object and, hence, its design plays a critical role in the tracker's overall success. Generally, appearance models are classified as generative and discriminative models. Generative models consider only foreground information but ignore background information for target representation. Generative models have been exploited for building trackers, such as distribution fields for tracking (DFT) [7] and an adaptive structural local sparse appearance-based tracker (ASLA) [8]. On the other hand, discriminative trackers separate the target from the background and consider tracking as a binary classification problem. Trackers, such as STRUCK [9] and the compressive tracker (CT) [10], were presented under the discriminative model. Discriminative trackers assume the background and foreground to have a linear relationship. However, this condition does not hold true in real-world scenarios. Generally, trackers based on the generative model tend to drift in a large search space, background clutters, occlusion, and large variations in target appearance, whereas trackers based on the discriminative model are more robust and reliable to these variations. However, the success of discriminative models is highly dependent on the selection of training samples and the optimal design of the updating

Manuscript received December 9, 2018; revised March 23, 2019; accepted May 21, 2019. This paper was recommended by Associate Editor Y. Yuan. *(Corresponding author: Kapil Sharma.)*

G. S. Walia is with SAG, Defence Research and Development Organisation, New Delhi 110054, India.

H. Ahuja, A. Kumar, and N. Bansal are with the Computer Science and Engineering Department, Delhi Technological University, New Delhi 110042, India.

K. Sharma is with the Information Technology Department, Delhi Technological University, New Delhi 110042, India (e-mail: kapil@ieee.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2019.2920289

2168-2267 © 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.

strategy. Alternatively, many deep-learning-based trackers have been proposed [11]–[13] to model the target's appearance. Although these models can automatically discover feature representation for the targeted object, they require extensive pretraining and are generally unemployable in real-time scenarios.

In this paper, we propose a real-time object tracking architecture based on unified graph fusion (UGF) and an adaptive appearance model which exploits the advantages of both discriminative and generative models. The discriminative appearance model is based on outlier detection and fragment distribution to tackle background clutter and occlusions. The UGF utilizes a proposed cross-diffusion-based generative process to build robust features invariant to illumination, scale, and deformations. To summarize, the main contributions of this paper are as follows.

1) We propose an object tracking framework based on the UGF of multiple features, which creates a highly marginalized decision boundary between feature descriptors of positive and negative fragments. This is achieved through the proposed cross-diffusion algorithm which captures the strong relationships between feature descriptors and suppresses any weak relations, making the tracker invariant to illumination and scale variations. The algorithm iteratively fuses sparse and dense similarity graphs and performs recursive normalization for faster convergence.
2) We propose an outlier detection mechanism based on the proximity of a candidate fragment to positive and negative class samples. The proximity measure evaluates the nodes of the decision trees in the random forest, where the candidate fragment lands. This ensures that the updating of the classifier does not compromise its accuracy. In addition, it enhances the selection of fragments for precise localization of the targeted object. This not only makes the tracker resistant to background clutter but also ensures efficient training of the classifier during initialization and updating of the appearance model.
3) We propose a novel adaptive appearance model to preserve temporal and spatial consistency and tackle deformations and pose variations. The mechanism distributes the proximal candidate fragments based on classifier confidence scores and sample reliability using multiclass SVM. The reliability metric prevents the appearance model from updating with fragments containing background clutters and ensures that the updating samples do not overfit the classifier. Using regular updates for a robust and dynamic appearance model and periodic updates of reference templates prevent eventual drift in long video sequences.

The remainder of this paper is organized as follows. In Section II, we closely review the literature to gauge the research gap. In Section III, we present the architecture of the proposed UGF tracker. In Section IV, we perform experimental validation and compare our tracker with 18 other state-of-the-art trackers over benchmarked video sequences. Finally, concluding remarks and future directions are sketched in Section V.

II. Related Work

Visual tracking has been rigorously investigated over the last decade. Generally, the performance of any visual tracking solution depends on building a robust and dynamic appearance model. In this section, we examine closely related object appearance models and their adaptation strategies to the dynamic environment.

Mostly, generative models represent the target's appearance using foreground information. In this direction, many sparse-based methods [14]–[16] were proposed. Li *et al.* [14] proposed weighted local sparse representation to measure the importance of each patch. The adaptive template update strategy considered incremental subspace and sparse coding for each reliable patch. On the other hand, in [16], appearance characteristics of the local patch were categorized as stable, invalid, and valid with importance weighting. Discriminative local sparse coding was used to separate the background patches from stable ones and a linear regressor was used to separate the invalid patches from valid patches. Wang *et al.* [3] used optical flow along with fusion of depth and color statistics for a two-stage object localization process, while Xue *et al.* [2] exploited three types of spatial–temporal context information for long-, medium-, and short-term components during visual tracking. Chen *et al.* [15] used frame difference masks to encode large-scale corruption information into a modulated dynamic sample dictionary. Qi *et al.* [17] exploited structure-aware local sparse coding for encoding each candidate sample. In this, encoding samples were reduced by using the global sparsity constraint on the coding coefficient. Generally, sparse-based generative methods cannot handle many of the tracking challenges due to a lack of background discriminative information.

On the other hand, discriminative appearance models separate the target from the background. In this direction, Huang *et al.* [18] proposed a part-based discriminative model wherein object appearance was constructed using a supervised descent method with two online reliabilities. For catering appearance variations, the updating model was made adaptive through incremental cascaded support vector regression. Fan *et al.* [19] proposed integrating subspace information and intrinsic geometric structures among tasks using dual-graph regularized discriminative multitask low-rank learning, whereas Yuan *et al.* [20] exploited traffic group behavior to support their tracking-by-detection framework and cater appearance variations in crowded scenes and heavy occlusion. Du *et al.* [21] used association and geometric encoding functions to compute the confidence of part-sets in a hypergraph. A mode-seeking function selected the most-probable parts over a candidate set. Yuan *et al.* [22] integrated depth maps with a superpixel sparse-encoding framework to model the target's appearance. In addition, many correlation filter-based appearance models were also proposed [23]–[25]. Sui *et al.* [23] proposed an elastic net constraint with a correlation filter to handle the tracking challenges. In this, the peak strength metric was defined to improve the capability of the filter. In [25], instead of a metric, Du *et al.* used an iterative graph-seeking strategy which minimizes an energy function based on correlation filters. Hu *et al.* [24] proposed the manifold regularization in the correlation filter for unlabeled data and

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

sample space. The tracker in [6] jointly treats the visual and geometric properties within a single formulation, with the correlation filter at the coarse level and a color model to localize the object.

Recently, many deep-learning-based trackers [26]–[28] were proposed pertaining to their favorable performance to model the target's appearance. Wang *et al.* [26] considered the problem of visual tracking as a trajectory estimation task using convolutional and recurrent units. Large training data are required to predict trajectories, which are often unavailable. Gao *et al.* [13] exploited part context information and importance reliability to preserve the target's spatial information using deep regression. Qi *et al.* [12] used a Siamese network with an online adaptive hedge algorithm for the similarity measure between the target and weak experts based on the CNN layer. Similarly, Gao *et al.* [11] trained a Siamese neural network to construct a structural relationship between image patches. Mostly, deep-learning-based visual trackers require pretraining of the neural network on a large auxiliary image dataset. To overcome pretraining, Zhang and Suganthan [28] initialized random convolutional filters and only train the fully connected layer. However, most deep learning-based trackers cannot be used in real-time scenarios and require computationally intensive fine-tuning.

Review of the closely related literature indicates that a future research direction should consider the design of a robust appearance model using multiple features with an online adaptive updating strategy, regardless of whether it is generative or discriminative. The next section details the UGF's methodology.

III. Proposed Methodology

The proposed UGF tracker leverages both motion cues and visual descriptors for precise localization of the object in video sequence. To adapt to variations in the object environment, the proposed tracker uses a robust updating strategy encompassing a proximity-based outlier detector, a transductive reliability metric, and classifier confidence scores. The appearance model of the object is based on the proposed cross-diffusion of HOG (F_e), local binary pattern (LBP) (F_t), and intensity (F_c) feature descriptors. The direct fusion of features is vulnerable to small changes in object appearance; hence, we use feature scores to calculate the fused feature descriptors of the given fragments. Fig. 1 illustrates the appearance model, object localization, and updating strategy of the UGF tracker. At $t = 1$, the architecture is initialized by sampling positive (S^+) and negative (S^-) fragments around the target object. This is followed by construction of unified feature descriptors $\mathrm{UF}_T(C^{(k)})$ from dense (U_e^k, U_c^k, U_t^k) and sparse (V_e^k, V_c^k, V_t^k) adjacency matrices of each candidate fragment $C^{(k)}$. These fragments are fed to a random forest classifier ($\mathcal{R}$) for training and are stored for construction of fused descriptors of candidate fragments during the prediction stage. The object localization during the prediction stage ($t = 2, 3, \ldots, N$) is a two-fold process. At each time step t, an initial rough localization using optical flow $\tilde{G}_t$ provides a more effective sampling state than G_{t-1} for feature extraction. Multiple extracted features of each candidate fragment are then fused using graph unification. The fused feature descriptors are fed to a two-phase classifier, which first classifies the positive and negative fragments followed by a secondary outlier detector which finds borderline positive and negative predictions $\hat{C}$ using a class proximity measure based on a leaf node count. The final target location G_t is obtained using the softmax probability of the positive fragment centroids. Further, at each time step, a transductive reliability metric ($r(\hat{C}^{(k)})$ is evaluated for each fragment. The reliability, along with the confidence score of proximal fragments is used for fragment distribution based on multiclass SVM. The positive ($\hat{C}^P$) and negative ($\hat{C}^N$) reliable fragments from this distribution are then fed to the classifier. The reference dictionary is also updated by selective replacement with positive and negative reliable fragments. This ensures that the classifier temporally adapts to the changes in object appearance.

A. Stage I: Rough Localization

In the first stage of object localization, the tracker evaluates motion cues from the apparent movement of pixel intensity values between consecutive frames. In the proposed architecture, we utilize the Horn–Schunk optical-flow method [29] for rough estimation of the target centroid in the subsequent frame. The method estimates the change in pixel intensity $\mathcal{I}(x, y, t)$ at time t, by assuming that the illumination variation between consecutive frames is zero. Using this assumption, we obtain (1) wherein the spatial gradient of intensity is represented by $\mathcal{I}_x$ and $\mathcal{I}_y$. $\mathcal{I}_t$ represents the change in intensity with respect to time, and u and v represent the estimation of the velocity field

$$\mathcal{I}_x u + \mathcal{I}_y v + \mathcal{I}_t = 0. \tag{1}$$

Assuming the smoothness constraint, velocities u and v are derived by minimizing the total error ($\mathcal{E}_T$) using

$$\mathcal{E}_T = \int\int \left(\mathcal{I}_x u + \mathcal{I}_y v + \mathcal{I}_t\right)^2 + \epsilon^2 \left(\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2 + \left(\frac{\partial v}{\partial x}\right)^2 + \left(\frac{\partial v}{\partial y}\right)^2\right) dxdy \tag{2}$$

where ϵ^2 scales the smoothness constraint. We select N optical-flow vectors in the region surrounding the target centroid G_{t-1}. The roughly estimated centroid $\tilde{G}_t$ in the current frame is then given by

$$\tilde{G}_{t,(x,y)} = G_{t-1,(x,y)} + \frac{1}{N}\sum_{i=1}^{N}\left(u^{(i)}, v^{(i)}\right). \tag{3}$$

This $\tilde{G}_t$ is used for generating candidate fragments in the current frame for precise localization of the target centroid.

B. Stage II: Precise Localization

The precise localization of the target is achieved via a multicue-based recursive graph fusion and normalization technique. First, we sample the fragments $C^{(k)} \in \{C^{(1)}, C^{(2)}, \ldots, C^{(K)}\}$ through a random walk around $\tilde{G}_t$. Each random walk is associated with the distance from $\tilde{G}_t$, a scale factor, and an angular displacement. For each candidate fragment, three complementary features are extracted and fused using the proposed cross-diffusion method.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

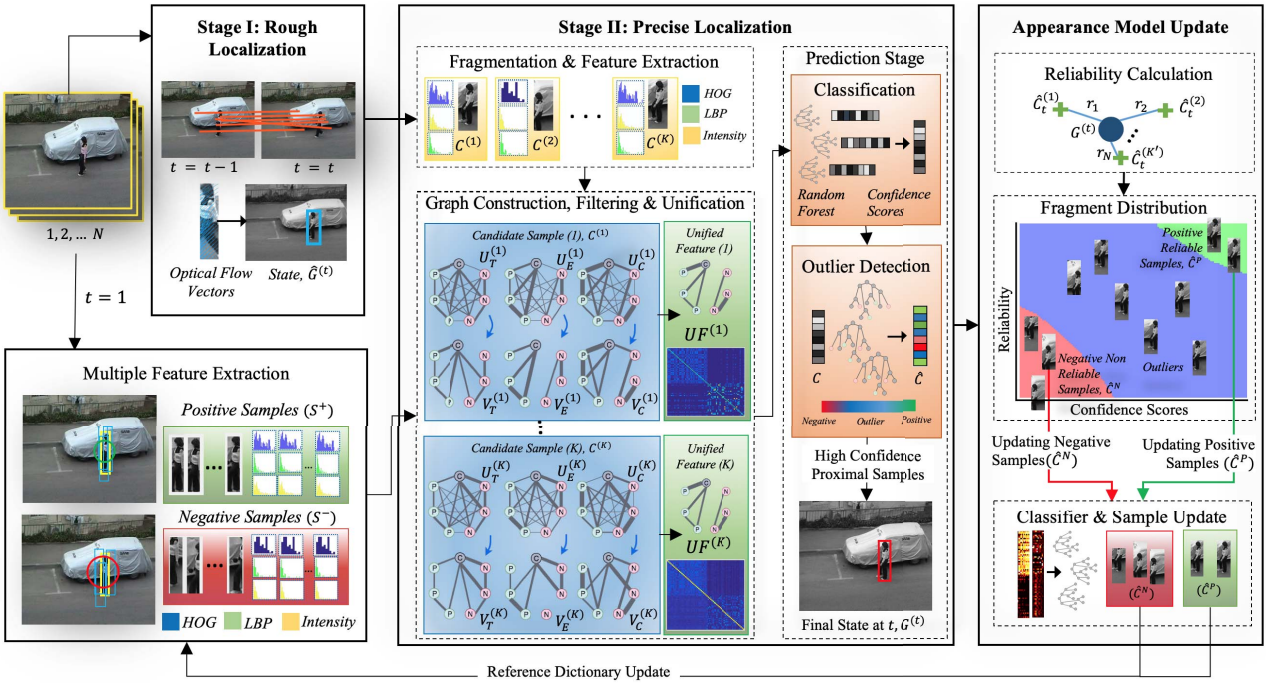


Fig. 1. Overview of the UGF tracking algorithm. The model is initialized ($t = 1$) by generating unified features for positive and negative fragments based on the proposed cross-diffusion algorithm. For each prediction step, $t = (2, 3, \ldots, N)$, the appearance model adapts to the current object environment based on an outlier detection mechanism and fragment distribution strategy considering classifier confidence scores and a transductive reliability metric.

1) Feature Extraction: The appearance characteristics of each candidate fragment $C^{(k)}$ are described by its HOGs, LBPs, and RGB histogram (intensity) feature descriptors. We use these feature descriptors due to their complementary nature which is exploited using the proposed method. The intensity histogram requires less computational load, is invariant to scale, and robust to partial occlusions. The intensity histogram of the kth candidate fragment is defined as $F_c^k = \{H_1, H_2, \ldots, H_{N_b}\}$, where the binning function assigns each pixel (x_k, y_k) to one of the N_b bins.

LBP feature descriptors are invariant to rotations, handling any inplane and out-of-plane rotations. The LBP operator is determined using

$$\mathrm{LBP}_{\beta,N_p} = \begin{cases} \sum_{j=0}^{N_p-1} u_\epsilon\left(\mathcal{I}_g^j - \mathcal{I}_g^c\right), & U\left(\mathrm{LBP}_{\beta,N_p}\right) \le 2 \\ p+1, & \text{otherwise.} \end{cases} \quad (4)$$

The texture uniformity $U(\mathrm{LBP}_{\beta,N_p})$ [30] measures the number of bitwise transitions of pattern labels. $\mathcal{I}_g^j$ and $\mathcal{I}_g^c$ represent the gray value of the central pixel (c) and the jth pixel, respectively. β is the neighborhood radius of the central pixel with N_p equally spaced neighboring pixels. For each candidate fragment $C^{(k)}$, normalized histograms are determined by mapping the uniformity measure for each pixel as F_t^k.

To describe the shape structure of the candidate fragments, we use the HOG feature descriptors [31] which capture the distribution of edge directions. The target object is divided into image patches or cells with each cell holding the weighted local 1-D histogram of gradient vectors over the pixels of the cell. A pixel votes for its orientation bin as well as its neighboring orientation bins. The bins obtained are $L2$-normalized for better invariance to shading and illumination. The histogram of orientation bins for the kth candidate fragment is stored as F_e^k as an edge feature cue.

2) Feature Unification: Our novel graph unification strategy performs nonlinear feature fusion with recursive normalization. This retains a strong representation of the target object and discards any weak features that make the tracker susceptible to changes in the object environment. A set of positive (S^+) and negative fragments (S^-) is stored during architecture initialization ($t = 1$) and updated with reliable fragments in each subsequent frame. During the feature unification procedure, the color (RGB), edge (HOG), and texture (LBP) feature descriptors for each candidate fragment $C^{(k)}$ are diffused into a single robust feature. For each feature descriptor $F_\mu^x \in \{F_t^x, F_e^x, F_c^x\}$ of fragment vertex $x \in N = \{C^{(k)}, S^+, S^-\}$, we construct graphs $U_\mu \in \{U_t, U_e, U_c\}$ with an edge weight defined as the likelihood between feature descriptors of two fragment vertices. Between any two fragments $x, y \in N$, we construct the similarity matrix $U_\mu \in \mathbb{R}^{N\times N}$ by calculating the Bhattacharya coefficient from their feature descriptors

$$\beta_\mu(x, y) = \sqrt{F_\mu(x) \odot F_\mu(y)}. \quad (5)$$

The Bhattacharya coefficient is used to calculate the Bhattacharya distance between the features of the fragments x and y as follows:

$$B_\mu(x, y) = \sqrt{1 - \beta_\mu(x, y)}. \quad (6)$$

We use the likelihood estimation on the Bhattacharya distance to calculate the likelihood between two feature descriptors as

$$U_\mu(x, y) = \frac{1}{\sigma_\mu\sqrt{2\pi}} e^{-\frac{(B_\mu(x,y))^2}{2\sigma_\mu^2}} \quad (7)$$

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

where $\sigma_\mu \in \{\sigma_t, \sigma_e, \sigma_c\}$ represents the cue standard deviation. The adjacency matrix U_μ is formed by calculating the likelihood between each of the N vertices of the graph. This is followed by row-wise normalization of U_μ, (8). The self-similarity is set to a small constant, ϵ for better distribution of weights among other similar fragments. The first row of the normalized similarity matrix $\hat{U}_\mu$ contains the edge weights corresponding to the candidate fragment $C^{(k)}$

$$\hat{U}_\mu(x, y) = \begin{cases} (1-\epsilon)\frac{U_\mu(x,y)}{\sum_y^N U_\mu(x,y)}, & x \neq y \\ \epsilon, & x = y. \end{cases} \tag{8}$$

From $\hat{U}_\mu$, we derive a sparse vector representation of the candidate fragment $V_\mu^k \in \mathbb{R}^{1\times N}$, which only keeps the most similar elements and sets the other elements to zero

$$V_\mu^k = \begin{cases} \hat{U}_\mu\big(C^{(k)}, y\big), & \text{if} \quad \hat{U}_\mu(y) \in k\text{-NN}\big(C^{(k)}\big) \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

In (9), $\hat{U}_\mu(y) \in \mathbb{R}^{1\times N}$ represents the row corresponding to adjacency list of the yth sample

$$\hat{V}_\mu^k = \frac{V_\mu^k}{\sum_{i=1}^N \big|V_\mu^k\big|}. \tag{10}$$

We normalize the derived vector representation $\hat{V}_\mu^k$, (10). This eliminates weak positive and negative relationships from the adjacency list of the candidate fragments. The normalization distributes the weights among the strong connections and, hence, builds a robust sparse representation. To fuse the individual feature sparse vectors $\hat{V}_\mu^k$, into a single robust feature descriptor $\text{UF}_T(C^{(k)}) \in \mathbb{R}^{1\times N}$, we propose a cross-diffusion strategy adapted from [32] for $M = |F_\mu|$ feature descriptors as

$$U_{\mu,t+1}\Big(C^{(k)}\Big) = \hat{V}_{\mu,t}^k \times \left(\frac{1}{M-1}\sum_{j=1, j\neq\mu}^{M} \hat{U}_{j,t}\Big(C^{(k)}\Big)\right) \times \hat{V}_{\mu,t}^k. \tag{11}$$

To increase computational efficiency of the diffusion process, we modified the iterative process by normalizing the row corresponding to the candidate sample in the similarity matrix $U_{\mu,t+1}(C^{(k)}) \in \mathbb{R}^{1\times N}$ obtained after each iteration as

$$\hat{U}_{\mu,t+1}\Big(C^{(k)}\Big) = \frac{U_{\mu,t+1}\big(C^{(k)}\big)}{\sum_{i=1}^N \big|U_{\mu,t+1}\big(C^{(k)}\big)\big|}. \tag{12}$$

Based on this new normalized adjacency list of $C^{(k)}$, we find its k-nearest neighbors from the original similarity matrix (9). Once the weak connections are eliminated, we normalize the sparse vector $V_{\mu,t+1}^k$ to obtain the normalized sparse vector $\hat{V}_{\mu,t+1}^k$ for the next iteration. Finally, the adjacency list of the candidate sample for each feature descriptor $\hat{U}_{\mu,t}(C^{(k)})$ is averaged to obtain the unified feature $\text{UF}_T(C^{(k)})$ by (13), which will be used as an input to the classifier for prediction

$$\text{UF}_T\Big(C^{(k)}\Big) = \frac{1}{M}\sum_{j=1}^{M} \hat{U}_{\mu,T}\Big(C^{(k)}\Big). \tag{13}$$

Here, T represents the final iteration of the cross-diffusion normalization process. Fig. 2 illustrates the efficiency of our algorithm against previously developed cross-diffusion algorithms. Each row represents the unified feature of the candidate fragment $\text{UF}_T(C^{(k)})$. The first 50 columns describe the similarity scores with positive fragments (S^+) and the next 50 columns represent the similarity scores with negative fragments (S^-). Further, for tracker validation, we take the first 200 candidate fragments to be positive and the next 200 fragments to be negative. The cross-diffusion algorithm [32] in Fig. 2(c) produces almost similar unified features for both negative and positive candidate fragments, making training of the classifier inefficient. This is due to the unstable self-similarity definition and the lack of a normalization procedure during iterative cross-diffusion. An improvement was proposed in [33] by normalization of the dense similarity matrix at each iteration. In Fig. 2(b), a high self-similarity constant in $U_\mu(C^k)$ reduces the distribution scores among other fragments, producing identical unified features for the positive candidate fragments. This method also produces anomalies, by describing negative candidates as positive. In the proposed method, the introduction of recursive normalization of the dense similarity matrix and vector sparse representation with an optimized self-similarity score leads to better distribution of similarity scores among other fragments. The recursive formation of $\hat{V}_\mu^k$ from $U_{\mu,t}(C^{(k)})$ highlights strong connections and diminishes the weaker connections unlike [32] and [33] which use the same sparse matrix in each iteration. Algorithm 1 summarized the procedure of the proposed cross-diffusion method.

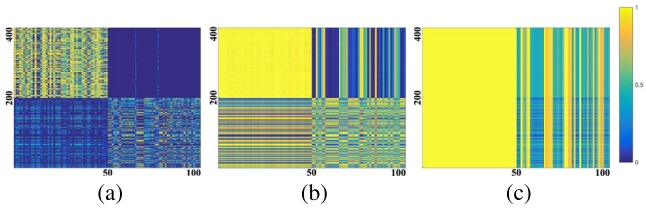


Fig. 2. Feature descriptors constructed using Bhattacharya's similarity metric. (a) Proposed cross-diffusion. (b) Modified cross-diffusion [33]. (c) Cross-diffusion [32].

3) Prediction Stage: The evaluation of prediction for precise localization is a two-fold process. At each time step $t = (2, 3, \ldots, N)$, the random forest classifier ($\mathcal{R}$) assigns each input feature vector $\text{UF}(C^{(k)})$ of candidate sample $C^{(k)}$, a prediction score. We adopted the bagging algorithm from Breiman *et al.* [34] which creates a single strong classifier from many weak classifiers. Each candidate feature is assigned a value in $[0, 1]$ according to its resemblance to the positive and negative fragments encountered during training of the classifier. At the end of this stage, we get a set of K scores for frame t, $\Delta_t = \{\delta_1, \delta_2, \ldots, \delta_K\}$. Next, we evaluate the proximity of the candidate sample $C^{(k)}$ toward a class as the proportion of decision trees $\mathcal{R}_t \in \mathcal{R} = \{\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_T\}$ for which $C^{(k)}$ falls on the same leaf node ($\mathcal{R}_t^l$) versus other candidate fragments that fall in the same class (14). For a particular class, proximity (p) is defined by (15)

$$\hat{p}\Big(C^{(k)}, C^{(j)}, \mathcal{R}_t\Big) = \begin{cases} 1, & \text{if } \mathcal{R}_t^l\big(C^{(k)}\big) = \mathcal{R}_t^l\big(C^{(j)}\big) \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

$$p\Big(C^{(k)}\Big) = \frac{1}{T}\sum_{t=1}^{T}\sum_{k=1, k\neq j}^{K} \hat{p}\Big(C^{(k)}, C^{(j)}, \mathcal{R}_t\Big). \tag{15}$$

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Algorithm 1: Proposed Cross-Diffusion

Function `GraphFusion`(C, S^+, S^-, σ)**:**
  for $(C^{(k)} \in C)$ **do**
    $N \leftarrow [C^{(k)}, S^+, S^-]$
    for $(F_\mu \in \{F_t, F_c, F_e\})$ **do**
      derive U_μ from Eq. (7) on N
      normalize U_μ to $\hat{U}_\mu$ using Eq. (8)
      if $\hat{U}_\mu(y) \in k\text{-}NN(C^{(k)})$ **then**
        $V_\mu^k \leftarrow \hat{U}_\mu(C^{(k)}, y)$
      else
        $V_\mu^k = 0$
      end
      normalize V_μ^k to $\hat{V}_\mu^k$ using Eq. (10)
      repeat
        find $U_{\mu,t+1}(C^{(k)})$ using $\hat{U}_{\mu,t}(C^{(k)})$ and $\hat{V}_{\mu,t}^k$ from Eq. (11)
        normalize $U_{\mu,t+1}(C^{(k)})$ to $\hat{U}_{\mu,t+1}(C^{(k)})$ using Eq. (12)
        $\hat{U}_{\mu,t}(C^{(k)}) := \hat{U}_{\mu,t+1}(C^{(k)})$
      until *convergence*
    end
    find $\mathrm{UF}_T(C^{(k)})$ using $\hat{U}_{\mu,T}(C^{(k)})$ from Eq. (13)
  end
  return $\{UF_T(C^{(1)}), UF_T(C^{(2)}), ..., UF_T(C^{(K)})\}$

Any $C^{(k)}$ with low proximity to their predicted sample class are considered outliers and are discarded to obtain the proximal candidate dictionary $\hat{C} = \{\hat{C}^{(1)}, \hat{C}^{(2)}, \ldots, \hat{C}^{(K')}\}$. The high confidence, proximal candidate fragments are averaged by the softmax probability of their confidence scores (16) to get the new precisely located centroid G_t at time step t (17)

$$\alpha_k = \frac{e^{\delta_k}}{\sum_{j=1}^{K'} e^{\delta_j}} \tag{16}$$

$$G_t = \sum_{k=1}^{K'} \alpha_k \hat{C}^{(k)}(x, y) \tag{17}$$

where $\hat{C}^{(k)}(x, y)$ represents the centroid of the kth candidate sample. Once the target object is localized, we update the appearance model of the object environment for consistent evaluation of subsequent frames.

C. Updating Strategy

To update the classifier $\mathcal{R}$ at each time step, we consider the proximal, high-confidence, and low-confidence fragments and calculate their reliability. The random forest classifier contains a bag of regression trees. Each its decision tree votes on the candidate fragment based on branching structure. The fact that a simple bag of regression trees is able to give high accuracy can be leveraged to distinguishing robust unified feature vectors. Further, the reliability metric $r(\hat{C}^{(k)})$ is used as a label of confidence for each selected fragment as

$$r\left(\hat{C}^{(k)}\right) = a \times \left(\tanh(-\left\|\hat{C}^{(k)} - G_t\right\|_2 + b) + c\right). \tag{18}$$

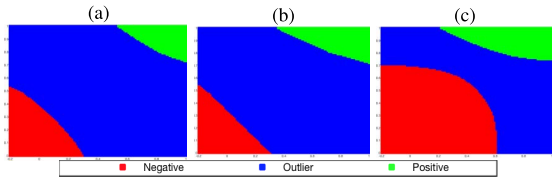


Fig. 3. Various kernels on multiclass SVM classification for fragment distribution. (a) RBF. (b) Linear. (c) Polynomial.

Here, a, b, and c are constants that describe the shape of the hyperbolic tan function. We derive a fragment distribution decision model, based on *one-versus-one* multiclass SVM, dividing the proximal candidate fragments into three categories: high confidence fragments with high reliability ($\hat{C}_P$); high confidence, low reliability, or low confidence, high reliability fragments labeled as outliers ($\hat{C}_O$); and low confidence, low reliability fragments ($\hat{C}_N$). Each data point, $\vec{x}_k = (r(\hat{C}^{(k)}), \delta_k)$ is represented as a pair of sample reliability and its confidence score. Each $\vec{x}_k$ is mapped to a $y_k \in \{\hat{C}_P, \hat{C}_N, \hat{C}_O\}$ for formulating the fragment distribution during initialization. For a pair of classes, $i, j \in \{\hat{C}_P, \hat{C}_N, \hat{C}_O\}, i \neq j$, $\vec{w}^{ij}$ is chosen to minimize the objective function as a binary classification problem for a penalty factor C and ξ_k^{ij} for relaxation from misclassified fragments as follows:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\left\|\vec{w}^{ij}\right\|^2 + C\sum_{k=1}^{K'} \xi_k^{ij} \\ \text{subject to} \quad & \begin{cases} \vec{w}^{ij} \cdot \vec{x}_k + b^{ij} \geq 1 - \xi_k^{ij}, & \text{if } y_k = i \\ \vec{w}^{ij} \cdot \vec{x}_k + b^{ij} \leq -1 + \xi_k^{ij}, & \text{if } y_k = j \end{cases} \\ & \xi_k^{ij} \geq 0. \end{aligned} \tag{19}$$

Using the *kernel trick* [35], we map the data to a nonlinear space for better generalization during predicton. The results for the linear, Gaussian, and polynomial kernel for sample classification are depicted in Fig. 3.

The RBF kernel $\mathcal{K}_{\mathrm{RBF}}$ (20) was able to generalize the best on the training data by generating optimal negative and positive class fragments and discarding most fragments as outliers

$$\mathcal{K}_{\mathrm{RBF}}(\vec{x}_i, \vec{x}_j) = e^{-\gamma\left\|\vec{x}_i - \vec{x}_j\right\|^2}. \tag{20}$$

This ensures a faster and efficient update of the random forest. The decision boundary is solved from a voting strategy where each of the M binary classifiers votes for a class on each incoming sample ($\hat{C}^k$). At each time step, the predicted $\hat{C}_P$ are used as positive samples for updating the classifier, whereas $\hat{C}_N$ from the proximal candidate fragments set $\hat{C}$ are used as negative samples. The selected candidate fragments, along with their respective labels, are fed to the random forest $\mathcal{R}$ for an update. To improve accuracy while tackling long-term visual tracking challenges, we also add a secondary periodic redetection procedure to identify the location of the target object when all candidate fragments are detected to be outliers, that is, when all sampled fragments are not proximal to either of the two classes. The redetection strategy utilizes a simple correlation filter around the previously estimated centroid (G_{t-1}) along with the estimated object trajectory to provide an estimation of the centroid in time step t. This prevents drift of

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Algorithm 2: Prediction and Updating Strategy

Function AppearanceModel($G_{t-1}, \mathcal{P}, \mathcal{R}, S^+, S^-$):
- $\tilde{G}_t \leftarrow$ HK-OpticalFlow(G_{t-1})
- $C \leftarrow$ Sampling($\tilde{G}_t, \mathcal{P}$)
- $\hat{C} \leftarrow \phi$
- **for** ($C^{(k)} \in C$) **do**
  - $\text{UF}_T \leftarrow$GraphFusion($C^{(k)}, S^+, S^-, \sigma$)
  - $\delta_k \leftarrow \mathcal{R}(\text{UF}_T(C^{(k)}))$
  - find proximity of $C^{(k)}$ using Eq. (15) & Eq. (14)
  - **if** $p(C^{(k)}) < P_{Threshold}$ **then**
    - discard $C^{(k)}$
  - **else**
    - $\hat{C}$.append($C^{(k)}$)
  - **end**
- **end**
- find α_k from $\Delta = \{\delta_1, \delta_2, ..., \delta_k\}$ using Eq. (16)
- find G_t using α_k from Eq. (17)
- **for** ($\hat{C}^{(k)} \in \hat{C}$) **do**
  - find reliability of $C^{(k)}$ using Eq. (18)
- **end**
- SVM $\leftarrow$ SVM Model ($r(\hat{C}), \Delta$)
- **for** ($\hat{C}^{(k)} \in \hat{C}$) **do**
  - $\{\hat{C}^P, \hat{C}^O, \hat{C}^N\} \leftarrow$ SVM($\hat{C}^{(k)}$)
- **end**
- Replace $\hat{C}^P$ in S^+ and $\hat{C}^N$ in S^-
- Update SVM Model with $\hat{C}^P$, $\hat{C}^O$ and $\hat{C}^N$
- Train $\mathcal{R}$ with S^+ and S^-
- Update sample set with $\hat{C}^P$, $\hat{C}^N$
- **return** $\mathcal{R}, G_t$

the tracker from the target object due to sudden changes in the object environment. Algorithm 2 describes the adaptation of the appearance model, along with the prediction of the target state in each frame. In the following section, we evaluate our tracker's performance against other state-of-the-art trackers.

IV. Experimental Validation

A. Experimental Setup

The appearance model of the proposed tracker is built using complementary multicue, namely, HOG, LBP, and RGB histograms. For fragment initialization, the sampling radius is defined to be large enough to capture the targeted object completely with minimal overlap of the fragments. The number of fragments chosen for initialization significantly depends on the dimensions of the targeted object, that is, a low-resolution target requires fewer samples than a target with high resolution. The tracking parameters used for simulation of the proposed tracker are defined as follows: the size of positive and negative sample dictionary are $(S^+) = 50$ and $(S^-) = 50$, respectively, the number of candidate sample generated in each frame is $|C| = 15$, and the reliability and prediction thresholds are set as $a = 0.5$, $b = 7$, $c = 0.7$, and $P_{\text{Threshold}} = 0.75$. To compare UGF with other tracking methods, we used MATLAB 2017b on an i7 2.2 GHz machine with 16 GB RAM. To run deep-learning-based trackers, we used an 8 GB GeForce GTX GPU.

TABLE I
Tracking Challenges and Video Sequences

Tracking challenges	Video Sequences
Occlusion & Background clutters (29 Sequences)	Football, Subway, MountainBike, Singer2, CarDark, Crossing, Car24, Bolt2, BasketBall, Walking, Jogging1, Jogging2, Walking2, Pedestrian1, Tiger, Soccer1, bike2, bike3, boat9, car11, car12, uav6, uav7, car15, person1_s, person3_s, person13, person18, person21
Scale variations (32 Sequences)	Car24, Walking, Dancer, Surfer, Walking2, Boy, Dog1, Crossing, Pedestrian1, Soccer1, Singer1, Shaking, bike2, bike3, boat5, boat7, boat9, building2, car4_s, car5, car11, car12, boat7, car15, person1_s, person3_s, person18, wakeboard2, uav6, uav7, person21, person23
Illumination Variations (15 Sequences)	Singer2, CarDark, Fish, Car24, BasketBall, Tiger, Singer1, Shaking, Soccer1, bike2, person3_s, car4_s, person1_s, person13, person21
Deformations & Fast motion (21 Sequences)	Walking, Subway, Jogging1, Jogging2, Dancer, Singer2, Crossing, Bolt2, Basketball, Pedestrian1, Soccer1, Bolt1, car4_s, boat7, person23, person1_s, wakeboard2, person3_s, wakeboard2, uav6, uav7,
In-plane Rotation (10 Sequences)	Surfer, Boy, Singer2, Football, MountainBike, Dancer, Dog1,Shaking, Bolt, Tiger
Out-of-plane Rotation (15 Sequences)	Surfer, Boy, Crossing, Singer2, Football, MountainBike, Dancer, Dog1, Basketball, Jogging1, Jogging2, Pedestrian1, Singer1, Shaking, Bolt1

B. Datasets and Evaluated Trackers

The proposed UGF tracker is evaluated on benchmark sequences from OTB-50, OTB-100 [36], VOT2017/18 [37], and UAV123 [38] datasets for rigorous performance analysis on various visual tracking challenges. A total of 25 914 frames consisting of video sequences from both stationary and moving cameras are used for evaluation. Further details of the challenges and their respective video sequences are tabulated in Table I.

The performance of the tracker is compared against 18 other state-of-the-art trackers. The evaluated trackers are included as follows:

1) *Generative Trackers (Two Trackers):* DFT [7] and ASLA [8].
2) *Discriminative Trackers (Six Trackers):* Structured output tracker (Struck [9]), CT [10], weighted multiple instance learning (WMIL) [39], incremental discriminative color learning tracker (IDCT) [40], fast discriminative scale space tracking (fDSST) [41], and multiple experts using entropy minimization (MEEM) [42].
3) *Correlation Filter-Based Trackers (Seven Trackers):* Discriminative correlation filter with channel and spatial reliability (CSRT) [43], attentional correlation filter network for adaptive visual tracking (ACFN) [44], context-aware correlation filter tracking (DCF-CA) [45], kernelized correlation filter (KCF) [46], learning continuous convolution operators for visual tracking (CCOT) [47], Staple: complementary learners for real-time tracking (STAPLE) [48], and multitask correlation particle filter for robust object tracking (MCPF) [49].
4) *Hybrid Tracker (One Tracker):* Sparsity-based collaborative model-based tracker (SCM) [50].
5) *Deep-Learning-Based Trackers (Two Trackers):* ECO: efficient convolution operators for tracking (ECO) [51] and learning multidomain convolutional neural networks for visual tracking (MDNet) [52]. Also, MDNet is the

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

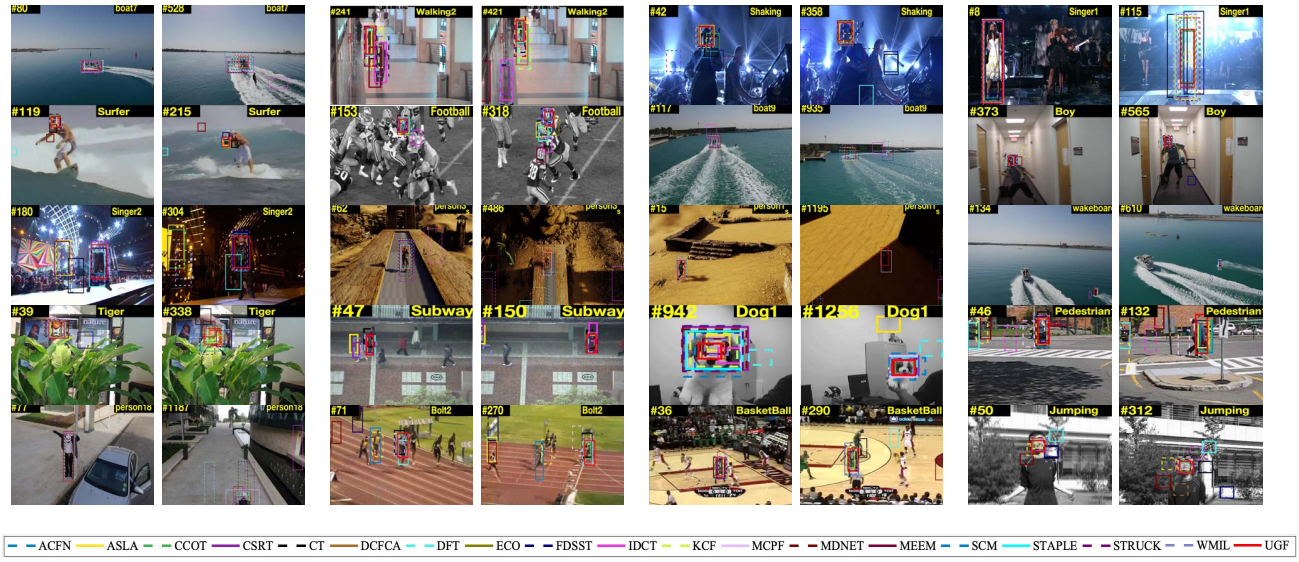


Fig. 4. Tracking results on critical frames for selected benchmark video sequences. The frame # and the name of the sequence are shown on the top left and right corner, respectively, of each frame.

winner of the VOT2015 challenge and CCOT is the winner of the VOT2016 challenge.

C. Evaluation Metrics

Seven performance metrics are considered for quantitative analysis in accordance with the VOT Toolkit [37]: center location error (CLE), *F*-Measure, distance threshold *versus* precision plot and success rate *versus* overlap threshold plot [36], accuracy, computational speed, and reliability. CLE is described as the Euclidean distance between target bounding box and the ground truth. ORE (Accuracy) is defined as $\Delta(\text{BB}_t \cap \text{BB}_g)/\Delta(\text{BB}_t \cup \text{BB}_g)$, where BB_t is the target bounding box and BB_g is the ground-truth bounding box. *F*-Measure is calculated as the harmonic mean of precision (P) and recall (R). Precision is defined as the overlap of BB_t and BB_g with respect to BB_t and recall determined by the overlap of BB_t and BB_g in reference to BB_g. The reliability is defined as the probability that a tracker will track the targeted object for up to S frames since failure and is calculated as $e^{-S^{F_o/N}}$. Here, N defines the total frames and F_o defines the total failures. Precision plots represent the mean precision at various center error thresholds (in pixels). The success plots illustrate the percent success over various overlap thresholds (in %). Overlap precision is the percentage of frames where bounding box overlap is above a threshold.

D. Challenge-Based Evaluation

In this section, we analyze and compare the performance of trackers under six tracking challenges for video sequence as described in Table I. Trackers output for challenging frames of the benchmark video sequences are illustrated in Fig. 4.

1) Object Deformation and Fast Motion: Generally, deformation in target appearance can occur due to pose variations or background clutter. In *Pedestrian1* sequence, there is a considerable change in the appearance of the targeted object, along with rapid camera movements. Ours, MDNet, ACFN, CCOT, and IDCT successfully keep track of targets throughout the whole sequence. The regular update of the tracker with reliable high confidence candidate fragments keeps track of the dynamic object environment. The generation of robust complementary features through the proposed cross-diffusion exploits LBP and RBG features over HOG since it deteriorates with deformations. *Bolt2*, *person3_s*, and *Tiger* sequences are full of pose variations. Most trackers fail as their appearance model is vulnerable to rapid change in object appearance except Ours, CCOT, MCPF, and ECO. In *Basketball* and *person1_s*, generative trackers, *viz.*, ASLA and DFT drifted from the target as they consider erroneous background pixels as the target. The *Jumping* sequence contains high object deformation, fast motion, and motion blur. The eventual erroneous update of the proposed tracker with blurred fragments results in comparatively lower performance to CCOT, but higher performance than other trackers due to a periodic update. UGF achieved an average precision score of 0.727 and an average success score of 0.606 [Fig. 5(a)].

2) Illumination Variations: Sudden and considerable illuminations variations are observed in *Singer2* and *person1_s*, where Ours and MCPF outperform other trackers. This is because the appearance model is built, with a considerably higher number of fragments compared to the incoming fragments. In *CarDark* and *person21*, the nonlinear multicue fusion of feature descriptors builds a robust representation of the object, making the tracker invariant to eventual changes in illuminations. Here, the HOG feature descriptor suppresses the inconsistencies in RGB and LBP cues, which are susceptible to change in illumination. In *CarDark*, ACFN, DFT, and CT fail to build a robust target representation under a low-contrast setting leading to eventual drift in these sequences. Our tracker shows comparatively low performance to SCM and CCOT on *Car24* due to fragment updates having minor background clutter after ≈1500 frames. The average precision score of 0.76 and average success score of 0.715 is attained by UGF as shown in Fig. 5(b), outperforming all other trackers.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

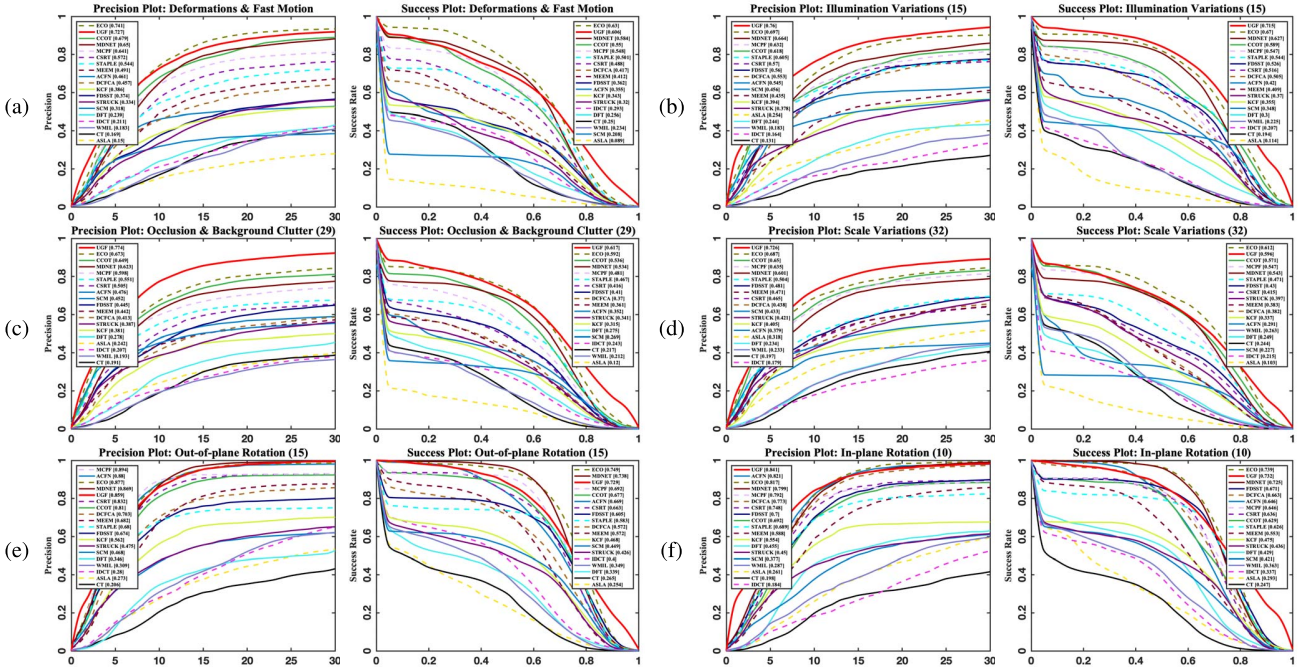


Fig. 5. Precision plots and success plots for (a) deformations and fast motion (b) illumination variations, (c) occlusion and background clutter, (d) scale variations, (e) out-of-plane rotation, and (f) inplane rotation. The legend ranks the trackers for each challenge, using the area under curve (AUC) for the success plot, and the score at 10 pixel threshold for the precision plot.

3) Occlusion and Background Clutters: In *MountainBike, Walking, car11*, and *boat9*, the target shows similar texture to that of the background. Ours with CSRT, CCOT, and MDNet is able to handle background clutter. The remaining trackers show minor drifts caused by eventual inconsistent updates. In *Football* sequence at frame #153, when the target is surrounded by other players, only the proposed approach, CCOT, STRUCK, ACFN, CSRT, and MDNet successfully keep track of the object, while DCFCA and MEEM are susceptible to clutter. This is due to the outlier detection mechanism which eliminates the nonreliable high-confidence fragments during an update, keeping the appearance model robust. Further, during partial occlusion of the targeted object, the inherent selective boosting of LBP and RGB features and suppressing HOG feature descriptors by the proposed cross-diffusion algorithm ensures precise localization. The target has been occluded by other people in subsequent frames (#47 and #150) of the *Subway* sequence. ASLA, WMIL, MEEM, and fDSST failed to recover under occlusion as there is no drift alleviation mechanism in them but Ours with CCOT, ACFN, and MDNet are able to cater to the drifts. Similarly, in *Jogging2*, when the target is occluded by the pillar, all but Ours, MDNet, CSRT, SCM, CCOT, and ACFN lose the targeted object. Under occlusion and background clutter, UGF outperformed the other trackers with an average precision score of 0.774 and an average success score of 0.617 [Fig. 5(c)].

4) Scale Variations: Although many trackers are able to localize the target object in *person18*, only Ours, STAPLE, KCF, DCFCA, and CCOT are able to accommodate scale variations. In *boat7* and *Walking2*, WMIL, ASLA, CT, DCFCA, MEEM, CSRT, and IDCT lose the target object to a similar occlusion due to inefficiency in handling variations in scale. At #421, Ours, CCOT, SCM, and ECO are efficiently able to adjust to the target's scale variation, whereas the appearance model in MEEM, KCF, ASLA, DFT, fDSST, and STRUCK gets corrupted by background clutter as the size of the targeted object is reduced. SCM and IDCT suffer in the sequences, *Dancer, Surfer, Boy*, and *Shaking* where scale variations are accompanied by inplane and out-of-plane rotations. Our method's periodic and regular updating strategy accompanied by complementary fusion of multicue accommodates various possible appearances of the object sampled at various scales, hence showing high performance compared to other trackers. Under this challenge, as shown in Fig. 5(d), the proposed tracker achieved the highest average precision score of 0.726 and highest average success score of 0.596.

5) Out-of-Plane Rotation: These occur in sequences where the target object moves out of the image plane. In *Dog1* and *Dancer*, and *Singer2*, out-of-plane rotation and scale variations are tackled via UGF by amplifying the LBP, and HOG feature descriptors over RGB. Under such settings, Ours, IDCT, MEEM, and KCF were able to tackle variations and efficiently track the target object. In *Singer*2, #304, CCOT, MEEM, ECO, and STAPLE among many other trackers fail to keep track of the target object. UGF handles out-of-plane rotations with background clutter in sequences *Football*, *MountainBike* and *BasketBall*, *Shaking* by consistently updating the online dictionary. Under this challenge, UGF achieved an average precision of 0.859 and an average success of 0.729 [Fig. 5(e)].

6) Inplane Rotation: This frequently occurs in image sequences due to the rotation of the target in the image plane. In *MountainBike* and *Surfer*, the target objects undergo a substantial amount of in-plane rotations throughout the sequence. UGF is able to handle eventual changes in orientation of the target object by regularly updating the sample dictionary. DFT, CT, IDCT, SCM, and STAPLE are not able to interpret the

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE II
FPS COMPARISON OF TOP-6 TRACKERS

Algorithm	MCPF	STAPLE	CCOT	ECO	MDNet	Ours
Accuracy	0.5355	0.5389	0.5722	0.6087	0.6107	**0.6639**
Speed (FPS)	0.570	**18.4**	0.480	6.1	1.14	15.3

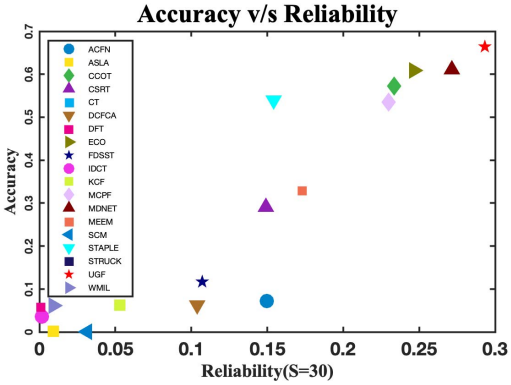


Fig. 6. Scatter plot for accuracy versus reliability ($S = 30$) on various trackers. The tracker placed in the top right corner corresponds to efficient performance, both in reliability and accuracy.

change in orientation of the target object and, hence, loose accuracy. In *Singer*1, *Boy*, *Dancer*, and *Dog*1, inplane rotations are coupled with scale variations which are tackled via the LBP feature descriptors in multicue feature extraction. In such scenarios, generative trackers fail due to erroneous updates of negative background samples. The proposed tracker shows an average precision score of 0.841 and average success score of 0.732 is attained by UGF [Fig. 5(f)].

E. Computational Complexity Analysis

The major computational load of the proposed tracker is in the cross-diffusion step where the sample dictionary is updated with positive reliable samples ($\hat{C}_P$), which involves multiple matrix multiplications and normalization operations (Algorithm 1: line 14). The proposed cross-diffusion reduces the complexity of the original cross-diffusion [32] from $\mathcal{O}(i \cdot N^4)$ to $\mathcal{O}(i \cdot N \cdot \hat{C}_P{}^2 + i \cdot N^2)$, where N is the size of the sample dictionary and i describes the number of iterations. Also, the number of iterations is reduced as the proposed cross-diffusion algorithm converges faster. To evaluate the computational demand of UGF, we have compared the proposed tracker against the other top-5 trackers in terms of speed (frames/s) as illustrated in Table II.

An efficient tracker needs to balance the tradeoff between accuracy, speed, and failure recovery. Fig. 6 illustrates the failure recovery probability of various trackers. UGF's high reliability is leveraged to its efficient updating methodology. ECO, STAPLE, and CCOT show competitive performance to UGF considering the tradeoff between computational time and accuracy (Fig. 7). Although MDNet performance falls close to UGF, the proposed tracker is ≈90% faster than MDNet. Hence, it achieves better accuracy with favorable computational demand against state-of-the-art trackers.

F. Overall Performance

The proposed tracker outperforms the compared state-of-the-art trackers MDNet, ECO, CCOT, STAPLE, and MCPF

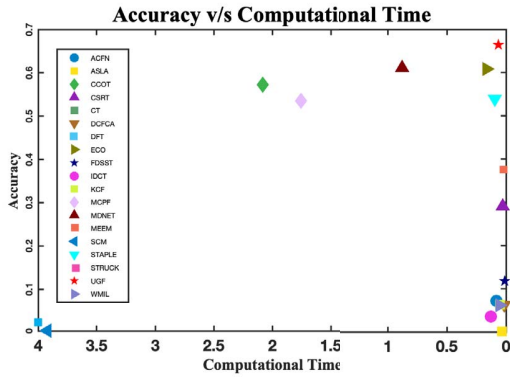


Fig. 7. Scatter plot for accuracy versus computational time (seconds) on various trackers. Tracker placed on the top right corner corresponds to efficient performance, both in computational time and accuracy.

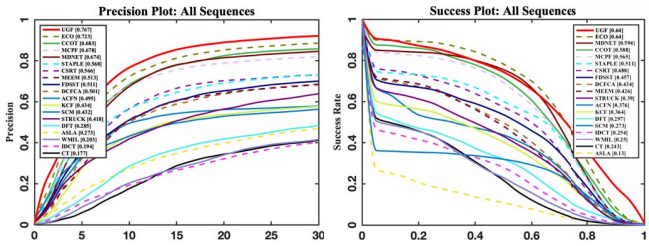


Fig. 8. Overall performance comparison of evaluated trackers: precision plots and success plots. The legend ranks the trackers, using the AUC for the success plot, and the score at 10 px threshold for the precision plot.

by 8.01%, 8.3%, 13.81%, 18.82%, and 19.34%, respectively, in terms of accuracy.

Tables III and IV describe the mean CLE and F-Measure of the trackers, respectively. Our proposed tracker achieved an average CLE of 13.59 and an average F-Measure of 0.767. In sequences with background clutter, occlusion, and inplane and out-of plane rotation, the superior performance of our tracker illustrates the effectiveness of the multifold adaptive appearance model which regularly updates the object appearance. The high overlap during illumination and scale variations validates the robustness of the fused feature descriptor against variations in scale and minor changes in object environment. In object deformations and partial occlusion, the complementary cues are generated via the proposed cross-diffusion exploited effective visual cues over deteriorating ones to ensure high precision. The proposed method significantly improves over existing fragment-based and classification-based models due to the above discussed cross-diffusion strategy and an adaptive update model. Further, Fig. 8 illustrates the success and precision of the overall performance of the tracker.

The UGF tracker addresses the limitations of existing state-of-the-art trackers to a great extent. Generative trackers like ASLA, IVT, and DFT were not very robust against background clutters and shape information; and discriminative trackers like STAPLE, WMIL, CT, and STRUCK tracker were unable to tackle target rotational and deformation challenges. STAPLE was unable to handle the dynamic appearance variations of the target and DFT was not able to cater to large pose variations. The KCF tracker used only a single cue for

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE III
AVERAGE CLE (IN PIXELS). THE BEST RESULTS ARE SHOWN IN RED, BLUE, AND GREEN, RESPECTIVELY

Challenge	ASLA	IDCT	KCF	DFT	CT	SCM	STRUCK	WMIL	CCOT	MEEM	CSRT	FDSST	ACFN	DCFCA	ECO	MCPF	MDNet	STAPLE	Ours
Occlusion & Background clutters	129.61	154.38	121.16	129.80	121.65	107.00	106.19	149.52	60.14	114.47	89.00	103.46	97.36	102.65	53.11	66.74	30.80	78.32	18.67
Out-of-plane Rotation	82.70	47.98	72.14	56.85	91.02	69.85	72.29	71.85	17.69	22.36	17.04	51.74	5.92	22.92	6.18	16.38	5.53	33.72	6.12
Scale Variations	114.99	160.04	109.29	138.34	111.01	104.60	78.58	129.37	51.58	84.01	86.93	92.00	115.05	93.68	50.18	50.68	30.60	72.72	21.17
Illumination Variations	119.65	161.65	136.51	130.24	146.55	111.70	98.84	160.64	46.72	121.92	68.76	105.48	117.54	70.50	35.61	34.25	19.57	65.65	15.29
Deformations and Fast motion	164.74	145.94	132.19	124.09	130.36	160.49	102.63	148.58	30.45	74.47	52.88	122.84	111.27	89.56	22.04	42.49	18.82	55.91	13.45
In-Plane Rotation	80.27	61.06	104.81	52.12	104.76	73.53	75.19	79.46	25.14	27.80	24.06	47.50	6.84	7.79	6.53	23.89	6.82	22.22	6.82
Overall Performance	**115.33**	**121.84**	**112.69**	**105.24**	**117.56**	**104.53**	**88.95**	**123.24**	**38.62**	**74.17**	**56.44**	**87.17**	**75.66**	**64.52**	**28.94**	**39.07**	**18.69**	**54.76**	**13.59**

TABLE IV
AVERAGE F-MEASURE. THE BEST RESULTS ARE SHOWN IN RED, BLUE, AND GREEN, RESPECTIVELY

Challenge	ASLA	IDCT	KCF	DFT	CT	SCM	STRUCK	WMIL	CCOT	MEEM	CSRT	FDSST	ACFN	DCFCA	ECO	MCPF	MDNet	STAPLE	Ours
Occlusion & Background clutters	0.135	0.286	0.371	0.329	0.272	0.294	0.404	0.259	0.631	0.429	0.492	0.480	0.417	0.440	0.685	0.570	0.621	0.539	0.71
Out-of-plane Rotation	0.321	0.489	0.543	0.414	0.332	0.511	0.498	0.432	0.775	0.683	0.767	0.682	0.785	0.669	0.846	0.788	0.842	0.652	0.836
Scale Variations	0.121	0.264	0.407	0.306	0.316	0.242	0.478	0.331	0.664	0.471	0.492	0.506	0.353	0.467	0.701	0.644	0.631	0.549	0.682
Illumination Variations	0.141	0.254	0.419	0.358	0.239	0.377	0.432	0.284	0.677	0.483	0.598	0.607	0.498	0.598	0.759	0.646	0.722	0.623	0.797
Deformations and Fast motion	0.091	0.352	0.393	0.299	0.310	0.220	0.396	0.300	0.662	0.513	0.589	0.427	0.434	0.494	0.743	0.653	0.695	0.580	0.735
In-Plane Rotation	0.376	0.417	0.548	0.518	0.312	0.494	0.508	0.446	0.731	0.665	0.740	0.763	0.772	0.779	0.843	0.746	0.834	0.707	0.840
Overall Performance	**0.197**	**0.344**	**0.447**	**0.371**	**0.297**	**0.356**	**0.452**	**0.342**	**0.690**	**0.541**	**0.613**	**0.577**	**0.543**	**0.574**	**0.763**	**0.675**	**0.723**	**0.608**	**0.767**

tracking and was not able to overcome variations in scale, whereas FDSST and FDSST were unable to sustain deformations. MCPF, ECO, and CCOT showed resistance to background clutters and illumination variations but were unable to redetect the target once the tracker drifted. MDNet had performance almost at-par with UGF but was comparatively slower.

To summarize, UGF shows superior performance against state-of-the-art trackers in both success rate and precision threshold with highest average F-Measure, lowest average CLE, high reliability, and favorable computational demands on benchmark video sequences.

V. CONCLUSION

In this paper, we have proposed a UGF visual tracking framework based on multicue cross-diffusion involving the fusion of sparse and dense similarity graphs with recursive normalization. The proposed cross-diffusion process generates complementary feature descriptors to handle object deformations and partial occlusions. We have also introduced a robust outlier detection mechanism which is able to discard fragments with background clutter and, hence, ensure a clear decision boundary between positive and negative fragments. Moreover, an adaptive update methodology is proposed to efficiently handle variations in scale and illumination. Extensive qualitative and quantitative analyses on sequences from the OTB-50, OTB-100, VOT2017/18, and UAV123 benchmark show the effectiveness of the proposed method against state-of-the-art trackers. On average of the outcome, the proposed UGF tracker achieves 13.59 CLE, 0.767 F-Measure, a success rate of 0.64, a precision score of 0.767, and accuracy of 0.6639. UGF also exhibits a competitive computational time of 15.3 frames/s.

In the future, the outlier detection mechanism can be extended to have fuzzy boundaries for smoother discriminability in the proximity space, but this process is computationally expensive. Another potential extension can be made by exploiting spatial relations between consecutive frames to build a holistic representation of the targeted object. Further, UGF can be extended to a multimodal environment by utilizing depth features.

REFERENCES

[1] G. S. Walia and R. Kapoor, "Recent advances on multicue object tracking: A survey," *Artif. Intell. Rev.*, vol. 46, no. 1, pp. 1–39, 2016.

[2] W. Xue, C. Xu, and Z. Feng, "Robust visual tracking via multi-scale spatio-temporal context learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2849–2860, Oct. 2018.

[3] Q. Wang, J. Fang, and Y. Yuan, "Multi-cue based tracking," *Neurocomputing*, vol. 131, pp. 227–236, May 2014.

[4] J. Fang, H. Xu, Q. Wang, and T. Wu, "Online hash tracking with spatio-temporal saliency auxiliary," *Comput. Vis. Image Understand.*, vol. 160, pp. 57–72, Jul. 2017.

[5] X. Lan, A. J. Ma, P. C. Yuen, and R. Chellappa, "Joint sparse representation and robust feature-level fusion for multi-cue visual tracking," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5826–5841, Dec. 2015.

[6] A. Lukežič, L. Č. Zajc, and M. Kristan, "Deformable parts correlation filters for robust visual tracking," *IEEE Trans. Cybern.*, vol. 48, no. 6, pp. 1849–1861, Jun. 2018.

[7] L. Sevilla-Lara and E. Learned-Miller, "Distribution fields for tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2012, pp. 1910–1917.

[8] X. Jia, H. Lu, and M.-H. Yang, "Visual tracking via adaptive structural local sparse appearance model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2012, pp. 1822–1829.

[9] S. Hare *et al.*, "Struck: Structured output tracking with kernels," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2096–2109, Oct. 2016.

[10] K. Zhang, L. Zhang, and M.-H. Yang, "Real-time compressive tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 864–877.

[11] J. Gao, T. Zhang, X. Yang, and C. Xu, "Deep relative tracking," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1845–1858, Apr. 2017.

[12] Y. Qi *et al.*, "Hedging deep features for visual tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 5, pp. 1116–1130, May 2019.

[13] J. Gao, T. Zhang, X. Yang, and C. Xu, "P2T: Part-to-target tracking via deep regression learning," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 3074–3086, Jun. 2018.

[14] Z. Li, J. Zhang, K. Zhang, and Z. Li, "Visual tracking with weighted adaptive local sparse appearance model via spatio-temporal context learning," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4478–4489, Sep. 2018.

[15] Z. Chen, X. You, B. Zhong, J. Li, and D. Tao, "Dynamically modulated mask sparse tracking," *IEEE Trans. Cybern.*, vol. 47, no. 11, pp. 3706–3718, Nov. 2017.

[16] K. Nai, Z. Li, G. Li, and S. Wang, "Robust object tracking via local sparse appearance model," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 4958–4970, Oct. 2018.

[17] Y. Qi, L. Qin, J. Zhang, S. Zhang, Q. Huang, and M.-H. Yang, "Structure-aware local sparse coding for visual tracking," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3857–3869, Aug. 2018.

[18] L. Huang, B. Ma, J. Shen, H. He, L. Shao, and F. Porikli, "Visual tracking by sampling in part space," *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 5800–5810, Dec. 2017.

[19] B. Fan, Y. Cong, and Y. Tang, "Dual-graph regularized discriminative multitask tracker," *IEEE Trans. Multimedia*, vol. 20, no. 9, pp. 2303–2315, Sep. 2018.

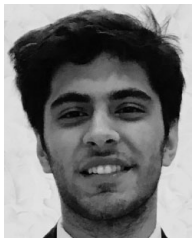
[20] Y. Yuan, Y. Lu, and Q. Wang, "Tracking as a whole: Multi-target tracking by modeling group behavior with sequential detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 12, pp. 3339–3349, Dec. 2017.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

[21] D. Du, H. Qi, L. Wen, Q. Tian, Q. Huang, and S. Lyu, "Geometric hypergraph learning for visual tracking," *IEEE Trans. Cybern.*, vol. 47, no. 12, pp. 4182–4195, Dec. 2017.
[22] Y. Yuan, J. Fang, and Q. Wang, "Robust superpixel tracking via depth fusion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 1, pp. 15–26, Jan. 2014.
[23] Y. Sui, G. Wang, and L. Zhang, "Correlation filter learning toward peak strength for visual tracking," *IEEE Trans. Cybern.*, vol. 48, no. 4, pp. 1290–1303, Apr. 2018.
[24] H. Hu, B. Ma, J. Shen, and L. Shao, "Manifold regularized correlation object tracking," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 5, pp. 1786–1795, May 2018.
[25] D. Du, L. Wen, H. Qi, Q. Huang, Q. Tian, and S. Lyu, "Iterative graph seeking for object tracking," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1809–1821, Apr. 2018.
[26] L. Wang, L. Zhang, and Z. Yi, "Trajectory predictor by using recurrent neural networks in visual tracking," *IEEE Trans. Cybern.*, vol. 47, no. 10, pp. 3172–3183, Oct. 2017.
[27] E. Gundogdu and A. A. Alatan, "Good features to correlate for visual tracking," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2526–2540, May 2018.
[28] L. Zhang and P. N. Suganthan, "Visual tracking with convolutional random vector functional link network," *IEEE Trans. Cybern.*, vol. 47, no. 10, pp. 3243–3253, Oct. 2017.
[29] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artif. Intell.*, vol. 17, nos. 1–3, pp. 185–203, 1981.
[30] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.
[31] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Vis. Pattern Recognit*, vol. 1, 2005, pp. 886–893.
[32] B. Wang, J. Jiang, W. Wang, Z.-H. Zhou, and Z. Tu, "Unsupervised metric fusion by cross diffusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2012, pp. 2997–3004.
[33] T. Tong, K. Gray, Q. Gao, L. Chen, and D. Rueckert, "Multi-modal classification of Alzheimer's disease using nonlinear graph fusion," *Pattern Recognit.*, vol. 63, pp. 171–181, Mar. 2017.
[34] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*. Boca Raton, FL, USA: CRC, 1984.
[35] B. Schölkopf and A. J. Smola, *Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2002.
[36] Y. Wu, J. Lim, and M.-H. Yang, "Object tracking benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1834–1848, Sep. 2015.
[37] M. Kristan *et al.*, "A novel performance evaluation methodology for single-target trackers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 11, pp. 2137–2155, Nov. 2016.
[38] M. Mueller, N. Smith, and B. Ghanem, "A benchmark and simulator for UAV tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 445–461.
[39] K. Zhang and H. Song, "Real-time visual tracking via online weighted multiple instance learning," *Pattern Recognit.*, vol. 46, no. 1, pp. 397–411, 2013.
[40] A. Asvadi, H. Mahdavinataj, M. R. Karami, and Y. Baleghi, "Online visual object tracking using incremental discriminative color learning," *CSI J. Comput. Sci. Eng.*, vol. 12, no. 2, pp. 16–28, 2014.
[41] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Discriminative scale space tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 8, pp. 1561–1575, Aug. 2017.
[42] J. Zhang, S. Ma, and S. Sclaroff, "Meem: Robust tracking via multiple experts using entropy minimization," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 188–203.
[43] A. Lukežič, T. Vojíř, L. Č. Zajc, J. Matas, and M. Kristan, "Discriminative correlation filter with channel and spatial reliability," in *Proc. CVPR*, vol. 6, pp. 4847–4856, 2017.
[44] J. Choi, H. J. Chang, S. Yun, T. Fischer, Y. Demiris, and J. Y. Choi, "Attentional correlation filter network for adaptive visual tracking," in *Proc. CVPR*, vol. 2, pp. 4828–4837, 2017.
[45] M. Mueller, N. Smith, and B. Ghanem, "Context-aware correlation filter tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, 2017, pp. 1387–1395.
[46] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, Mar. 2015.
[47] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 472–488.
[48] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P. H. S. Torr, "Staple: Complementary learners for real-time tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 1401–1409.
[49] T. Zhang, C. Xu, and M.-H. Yang, "Multi-task correlation particle filter for robust object tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4335–4343.
[50] W. Zhong, H. Lu, and M.-H. Yang, "Robust object tracking via sparsity-based collaborative model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2012, pp. 1838–1845.
[51] M. Danelljan, G. Bhat, F. Shahbaz Khan, and M. Felsberg, "ECO: Efficient convolution operators for tracking," in *Proc. CVPR*, 2017, pp. 6931–6939.
[52] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4293–4302.

Gurjit Singh Walia (SM'18) received the B.E. degree in electronics and communication from Guru Nanak Dev Engineering College, Ludhiana, India, in 1998, the M.E. degree in electronics from Punjab Engineering College, Chandigarh, India, in 2000, and the Ph.D. degree in computer vision from Delhi Technological University, New Delhi, India, in 2016.

He is a Senior Scientist with Defence Research and Development Organization, New Delhi. His current research interests include machine learning, pattern recognition, and information security.



Himanshu Ahuja (S'18) is currently pursuing the B.Tech. degree in computer science and engineering at Delhi Technological University, New Delhi, India.

His current research interests include biologically inspired models, object detection and tracking, graph-based knowledge systems, and sequential data modeling.

Ashish Kumar (M'16) is currently pursuing the Ph.D. degree in computer science and engineering from Delhi Technological University, New Delhi, India.

He has published many research papers in various reputed national and international journals and conferences. His current research interests include object tracking, image processing, video analysis, and machine learning.

Nipun Bansal (M'18) received the B.Tech. degree in information technology from Delhi Technological University, New Delhi, India, in 2011 and the M.Tech. degree in computer science from the Indraprastha Institute of Information Technology Delhi, New Delhi, in 2014.

He is an Assistant Professor with Delhi Technological University. He has executed a number of research projects in field of information security and computer vision. His current research interests include information security, image processing, and artificial intelligence.

Kapil Sharma (M'09) received the B.E. degree in computer science from Maharshi Dayanand University, Rohtak, India, in 2000, the M.Tech. degree in computer science from the Institute of Advanced Studies in Education, Sardarshahar, India, in 2005, and the Ph.D. degree in computer science from Maharshi Dayanand University in 2011.

He is a Professor and the Head of the Department of IT, Delhi Technological University, New Delhi, India. He has published over 100 research papers in referred journals. His current research interests include pattern recognition, computer vision, and soft computing.

The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)

What's Most Broken? A Tool to Assist Data-Driven Iterative Improvement of an Intelligent Tutoring System

Mononito Goswami*
Delhi Technological University
New Delhi, India
mononitog@hotmail.com
+91-8800592994

Shiven Mian*
IIIT-Delhi
New Delhi, India
shiven15094@iiitd.ac.in
+91-8130562083

Jack Mostow
Carnegie Mellon University
Pittsburgh, PA, USA
mostow@cs.cmu.edu

Abstract

Intelligent Tutoring Systems (ITS) have great potential to change the educational landscape by bringing scientifically tested one-to-one tutoring to remote and under-served areas. However, effective ITSs are too complex to perfect. Instead, a practical guiding principle for ITS development and improvement is to fix what's most broken. In this paper we present SPOT (Statistical Probe of Tutoring): a tool that mines data logged by an Intelligent Tutoring System to identify the 'hot spots' most detrimental to its efficiency and effectiveness in terms of its software reliability, usability, task difficulty, student engagement, and other criteria. SPOT uses heuristics and machine learning to discover, characterize, and prioritize such hot spots in order to focus ITS refinement on what matters most. We applied SPOT to data logged by RoboTutor, an ITS that teaches children basic reading, writing and arithmetic.

INTRODUCTION

An Intelligent Tutoring System (ITS) is a computer system that enables learning in an effective and meaningful manner by providing personalized instruction to learners. Prior research has demonstrated that ITSs take extensive time to author, with reported estimates of 200-300 hours of development per hour of instruction (Aleven et al. 2006). As a result, many authoring tools such as the Cognitive Tutor Authoring Tool (CTAT) have been built to make ITS development more efficient. But, despite our growing understanding of human cognition and the tutor authoring process, developing effective tutoring systems with limited development resources remains hard. In this situation, a practical guiding principle towards ITS development and improvement is to fix what is most broken.

The design of RoboTutor (RoboTutor 2015) was inspired by this principle. RoboTutor is an ITS developed as an open-source Android application that teaches children aged 7-10 in developing countries basic reading, writing, and arithmetic. In this paper, we address the question: is there a way to use data from RoboTutor to automate discovery of design issues and highlight them? To best address this issue, we must answer the following subquestions pertaining to the development and design process of RoboTutor:

- **Reliability:** How often and under what conditions does RoboTutor crash or hang? How fast does a child recover from the crash or hang?
- **Recognition**: How accurately does RoboTutor recognize both written and spoken input?
- **Usability:** How easily and efficiently can children operate RoboTutor? Which activities do they find hard to navigate?
- **Engagement:** When are the children disengaged the most, and why?

To answer these questions we present Statistical Probe of Tutoring (SPOT): an Educational Data Mining tool intended to help ITS developers identify 'what's most broken' – i.e, 'hot spots' with respect to design criteria such as software reliability, student engagement etc. SPOT uses metrics to identify and predict the occurrence of undesirable events, and trains a decision tree to discover hot spots. A hot spot is a subtree with a high proportion of undesirable events.

The intuitions that inspired this approach are as follows:

- Within a decision tree, undesirable events in the same subtree are likely to have the same underlying cause.
- The feature combination associated with a subtree - that is, the sequence of tests from the root to the subtree - characterizes when the undesirable events tend to occur and may reflect this underlying cause.
- Screen capture videos of a random sample of undesirable events in the subtree may shed further light on a certain cause and inspire ideas for how to address it.

Henry et al (Henry, Lin, and Park 2017) designed a UI/UX for SPOT and prototyped it using simulated data. Here we describe an implementation that uses automated decision tree learning and runs on real data from RoboTutor. We briefly describe the SPOT workflow and its findings on RoboTutor data in the next sections.

Methodology

Data Collection

RoboTutor uses a variety of activities such as story reading, math activities etc. to teach children. The dataset used by SPOT to discover hot spots is derived from student attempt

*authors contributed equally

Copyright © 2019, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

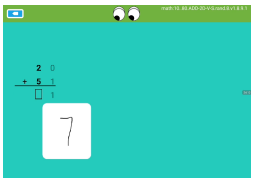


Figure 1: RoboTutor mis-recognizing 7 as 1 because of absence of middle stick - instance provided by SPOT

performance logs recorded from RoboTutor's beta field testing sites in Tanzania. We also extract activity and item-level aggregates from these performance logs. SPOT also uses screen capture videos of RoboTutor recorded using AZ Screen Recorder[1], used by developers to further examine the obtained hot spots.

Approach

To automate discovery of design issues, SPOT uses a work flow summarized as follows:

1. Developers specify a design criterion, and pick metric(s). A design criterion is any property of an Intelligent Tutor that we wish to analyze. We classify design issues into one of these design criteria: Reliability, Recognition, Usability and Engagement. Using pre-defined metrics (heuristics), SPOT approximately labels each row of the data as: suspicious instance indicating a design issue, or a non-suspicious instance. For example: the rejection rate of responses in writing activities is a metric of writing recognition, and a suspicious instance is any written response other than the expected answer.
2. SPOT trains a decision tree on the labeled data and discovers the top N hot spots using the F_1 score for each subtree. SPOT uses features such as activity name, student input, expected answer, attempt duration etc, with different features for attempt, item and activity-level data. The F_1 score enables SPOT to discover hot spots in the decision tree and balances both the concentration and the number of suspicious instances in the hot spots.
3. SPOT automatically characterizes each of the hot spots by conjoining every test on the path from the root node of the tree to the hot spot. In our work, we use Decision Trees strictly for finding and characterizing hot spots, rather than classifying unseen data.
4. SPOT picks up a random sample of suspicious instances of a hot spot, and presents their corresponding screen capture videos scrolled to the start of the instances along with the hot spot characterization, to the developers. Developers examine the videos and characterization for clues to diagnose the design issue, and devise solutions for redesigning RoboTutor.

Figure 1 illustrates a screen shot from the screen capture videos for a recognition hot spot characterized as: 'expected answer = 7'. After examining the sample of screen capture videos from this hot spot, we realized that the writing recognizer often mis-recognizes 7 in the absence of its middle stick as 1, leading children to confuse between the two.

[1]http://bit.ly/azurl

Conclusions

Table 1 presents some key SPOT findings and corresponding design implications for RoboTutor. We evaluated SPOT and selected features qualitatively, since design changes are fundamentally subjective decisions on part of the developers.

Criterion	SPOT Findings	Design Implications
Reliability	Counting and Story Reading actvities have high crash rates.	Implement crash logging, Examine Counting and Story Reading activities for bugs
Recognition	Children confuse between number pairs like 1 and 7, 5 and 3.	Bias data sources of appropriate activities towards frequently confused digits for better practice
Usability	Children spend unusually long time in Story Reading activities	Add timeouts to story reading activities
Engagement	Children bail out of activities when given the same problems repeatedly	Children should not be given the same problems repeatedly

Table 1: SPOT Findings and Implications

We believe SPOT would benefit many tutor developers by helping them iteratively improve their tutors and save time. SPOT may be especially useful when user-testing in person is impractical, such as in situations where the users are far away, when children may behave differently when adults are present, and when hot spots are important to fix but too rare to observe in person. SPOT can also be used for other ITSs, since it relies on simple data sources such as log files and screen record videos, both of which are easy to record.

Acknowledgements

Special thanks to other members of the RoboTutor team, especially Evelyn Yarzebinski for data parsing and cleaning. This work was sponsored by the Robotics Institute Summer Scholars (RISS) program, Carnegie Mellon University.

References

Aleven, V.; McLaren, B. M.; Sewall, J.; and Koedinger, K. R. 2006. The cognitive tutor authoring tools (ctat): Preliminary evaluation of efficiency gains. In *International Conference on Intelligent Tutoring Systems*, 61–70. Springer.

Henry, G.; Lin, J.; and Park, C. Y. 2017. SPOT: Refining robotutor. *HCI senior capstone project presentation, Carnegie Mellon University*.

RoboTutor. 2015. RoboTutor. http://robotutor.org.