A THESIS

ON

Big Data Analytics for Healthcare

BY

Ms. Amrita Sisodia

(2K17/PHD/CO/01)

UNDER THE SUPERVISION OF

Prof. Rajni Jindal

Professor

Department of Computer Science and Engineering

Delhi Technological University, Delhi

Submitted in partial fulfilment of the requirements of the

DOCTOR OF PHILOSOPHY
IN
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY, DELHI

INDIA

2022

DECLARATION

I, Amrita Sisodia a full-time Teaching cum Research Fellow (TRF) Roll No: 2K17/PHD/CO/01, hereby declare that the thesis entitled “**Big Data Analytics for Healthcare**” which is being submitted for the award of the degree of Doctor of Philosophy in Computer Science & Engineering, is a record of bonafide research work carried out by me in the Department of Computer Science & Engineering, Delhi Technological University, Delhi. I further declare that the work presented in the thesis has not been submitted to any university or institution for the award of any diploma or degree.

Amrita

Ms. Amrita Sisodia
Roll No: 2K17/PHD/CO/01
Department of Computer Science and Engineering
Delhi Technological University, Delhi

CERTIFICATE

This is to certify that the work embodied in the thesis entitled **“Big Data Analytics for Healthcare”** submitted by **Ms. Amrita Sisodia**, Roll No: 2K17/PHD/CO/01 as a full-time Teaching cum Research Fellow (TRF) in the Department of Computer Science & Engineering, Delhi Technological University, is an authentic work carried out by her under my guidance. This work is based on the original research and has not been submitted in full or in part for any other diploma or degree of any university to the best of my knowledge and belief.

Supervisor

Prof. Rajni Jindal
Computer Science & Engineering Department
Delhi Technological University, Delhi 110042

ACKNOWLEDGEMENT

I owe a debt of my gratitude to my supervisor Prof. Rajni Jindal, Department of Computer Science & Engineering, Delhi Technological University and for her valuable guidance, motivation, constant support and encouragement, which has helped me in formulating an urge for this research work. Her exemplary hard work and insightful guidance has always helped me in overcoming difficulties. I would like to thanks HOD of department of Computer Science and Engineering, SRC, DRC members, faculty members, and research scholars, staff of the department for their valuable support and time to time help. Finally, I would like to express my heartfelt regards to my family.

Ms. Amrita Sisodia
Roll No: 2K17/PHD/CO/01
Department of Computer Science & Engineering
Delhi Technological University
Delhi-110042

ABSTRACT

The rapid development of urbanization improved our daily life but also leads to a series of urban diseases. Worldwide around 2.4 million deaths could be prevented if a practice of proper hygiene can be maintained. Nowadays the focus of the government is to provide better and safe health to its citizens. So many new policies are coming into existence and a large number of funds have been assigned to implement these policies. The Indian government has recently announced the world's largest health care program under the Ayushman Bharat scheme. In this Program Policymakers and business practitioner around the world extending their extraordinary efforts in the field of e-health. Many local people awareness programs are also running which provides information to the citizens about infectious diseases like AIDS, Swine flu, Malaria, Covid-19, etc. These awareness programs run under the supervision of local health authorities. Since the digitization makes the work easier to spread awareness among the people. The primary requirement of a government is to provide a healthy environment and it can be achieved through digitalization.

The development of computer science and regional information generates a vast amount of data and creates an environment for disease prediction. Big data analytics (BDA) is a revolution in information technology and can be used for healthcare sector to perform data analysis. As the data is accumulated from various sources such as wearable sensor technology, smart phone, robots, IoT, etc. This data can be used for clinical predictions by using emerging machine learning algorithms and BDA techniques. BDA provides systematic information based on the vast amount of healthcare data to develop better healthcare system. With the help of the analyzed patterns, valuable information can be extracted and used by the policymakers to build a protective environment for the better healthcare system.

Therefore, the present research work integrates data management scheme for healthcare along with data analysis system to achieve the target towards health 4.0 for providing better healthcare service to the individual users. This research aim is to incorporate the five significant contributions to the healthcare industry.

First, the literature review of big data analysis has been performed for healthcare technology to highlight the challenges of the healthcare industry and identify the various mechanism to overcome these challenges.

Second, we proposed a novel healthcare multi-phase architecture (HCMP) to predict chronic kidney disease. The HCMP architecture works on six different layers namely: data-collection, data-storage, data-management, data-processing, data-analysis, and report-generation. The data-storage and data-management layers were performed on heterogeneous Hadoop cluster and the profiling methods were used to consider three situations for calculating the capacity ratio of each DataNode in the cluster. MapReduce is used for parallel data processing. Furthermore, horizontal scaling is performed in the Hadoop cluster, and the performance of every DataNode is investigated based on a capacity ratio. The data-analysis layer has performed classification tasks using a Decision tree, K Nearest Neighbors (KNN) classification, Kernel distributed Naïve Bayes, Simple distributed Naïve Bayes, Random Forest, and Random tree. Among these classifiers, Kernel distributed Naïve Bayes has produced the best results. The experiments are performed using tools such as Hadoop and RapidMiner to evaluate, analyze the efficiency and performance of the proposed architecture.

Third, the HCMP architecture also deployed the proposed MySymptom algorithm to filter the Chronic Kidney Disease (CKD) dataset of patients according to their symptoms at the data processing layer. The case study of CKD in Indian environment has been explored.

Forth, the proposed work is enhanced to analyze multiple disease using a new Hadoop-based Optimal Healthcare Classification Multi-Disease Diagnostic (OHCMDD) architecture for handling multiple diseases database with reduced features set. The proposed OHCMDD architecture also handles the DataNode deletion problem in Hadoop's cluster. Also, an intelligent classification prediction model is introduced for healthcare, namely Density-Based Features Selection with Spider Monkey Optimization (D-SMO) is used for chronic kidney disease (CKD), heart and diabetes disease. The empirical results of the D-SMO algorithm are compared with existing methods. The presented intelligent multi-disease model outperformed the other methods and implemented using the Hadoop cluster, HDFS and R platform.

Lastly, the proposed e-health architecture is deployed to achieve the target of health 4.0. The future of health management will become timelier and more personalized. As new technologies will empower individuals to conduct their health monitoring by using cyber–physical systems. The design principles of industry 4.0 connect the physical and virtual world in real-time.

Virtualization in health happens after the emergence of Information and Communication Technologies (ICT). For this 5G, the next-generation mobile network provides ambient intelligence for orchestration of medical services so that government and private companies can reconsider health prospects. These technological developments in healthcare, big data and industry 4.0 are individually helping to achieve the goal of health 4.0. The proposed work also provides the road map for health 4.0 along with different technologies involved in it.

The analysis and performance evaluation of the experimental results demonstrate that the proposed work provides a reliable architecture for better healthcare environment. Moreover, the comparative analysis displays that the proposed work showed an improvement. Thus, the proposed study provides an optimal and effective architecture for healthcare industry towards health 4.0.

Contents

Declaration……………………………………………………………………... i
Certificate……………………………………………………………………. ii
Acknowledgement…………………………………………………………… iii
Abstract……………………………………………………………………… iv
Contents……………………………………………………………………… v
List of Abbreviations………………………………………………………… xi
List of Figures……………………………………………………………….. xiii
List of Tables………………………………………………………………… xvi
List of Algorithms…………………………………………………………… xviii

CHAPTER 1: Introduction

1.1 Introduction………………………………………………………………… 1
1.1.1 Industry 4.0…………………………………………………………... 2
1.1.2 Health 4.0……………………………………………………………. 2
1.1.2.1 H4.0 Towards Improved Healthcare…………………………… 3
1.1.3 RAMI 4.0……………………………………………………………. 4
1.1.4 Big Data…………………………..……………………………… 5
1.1.5 Data Management…………………………………………………… 8
1.2 Motivation & Scope…….………………………………………………….. 9
1.3 Problem Statement………………………………………………………. 10
1.4 Research Objective……………………………………………………… 10
1.5 Contribution of the Thesis….…………………………………………………. 11
1.6 Thesis Organization………………………………………………………… 13

CHAPTER 2: Literature Review

2.1 Introduction………………………………………………………………… 16
2.2 Planning the Review……………………………………………………... 16
2.3 Conducting the Review…………………………………………………….. 17
2.4 Reporting the Review……………………………………………………. 18
2.4.1 Traditional Healthcare Data Handling Techniques and Classification Techniques for Healthcare…………………………………………….. 18
2.4.2 H4.0 Approach for Healthcare…………………………………………….. 22
2.4.3 Identification of Research Gaps………………………………………... 30
2.5 Summary……………..…………………………………………………….. 32

CHAPTER 3: Exploring Big Data for Healthcare

3.1 Introduction……….……………………………………………………….. 33
3.2 Inconsistent Data…..………………………………………………………... 34
3.2.1 Senior Citizens………………………………………………………….. 34

3.2.2 Citizens' Health………………………………………………………. 35
3.2.3 Increase in Chronic Disease………………………………………….. 35
3.2.4 Government Policy and Mandate…………………………………….. 35
3.3 Big Data Analysis and Associate Challenges………………………………... 36
3.3.1 Volume……..……………………………………………………….. 36
3.3.2 Velocity……..……………………………………………………….. 37
3.3.3 Veracity…………………………………………………………………… 37
3.3.4 Variety……………………………………………………………….. 37
3.3.5 Validity……………………………………………………………… 37
3.3.6 Volatility……………………………………………………………... 37
3.3.7 Value………………………………………………………………… 38
3.4 Big Data in Healthcare…..…………………………………………………….. 38
3.5 Big Data Analysis Using Different Tools……………………………………… 39
3.6 Healthcare Ecosystem……………………………………………………… 40
3.7 Experimental Method………………………………………………………… 42
3.7.1 Result and Discussion………………………………………………… 43
3.7.2 Comparison with Other Techniques…………………………………….. 46
3.8 Summary………………………………………………………………………… 46

CHAPTER 4: Healthcare Multi-Phase Architecture for Disease Prediction

4.1 Introduction……………………………………………………………………… 48
4.2 Proposed Work…………………………………………………………………... 51
4.2.1 HCMP Architecture…...…………………………………………………... 51
4.2.1.1 Data Collection Layer…………………………………………... 51
4.2.1.2 Data Storage Layer……………………………………………… 52
4.2.1.3 Data Management Layer……………………………………….. 53
4.2.1.4 Data Processing Layer………………………………………….. 57
4.2.1.5 Data Analysis Layer…………………………………………….. 61
4.2.1.6 Report Generation Layer……………………………………….. 62
4.3 Data Analysis…..……………………………………………………………... 62
4.3.1 Dataset and Attributes…………………………………………………... 62
4.4 Case Study: The Integrative Investigation of CKD…………………………… 64
4.5 Experimental Setup……..……………………………………………………... 65
4.6 Result and Discussion………………………………………………………… 66
4.6.1 Analysis Based on AUC-ROC Plot……………………………………… 67
4.6.2 Comparison with Other Proposed Techniques…………………………... 73
4.7 Summary………………………………………………………………………… 76

CHAPTER 5: D-SMO Based Multi Disease Prediction Model

5.1 Introduction……………………………………………………………………… 78

5.2 Proposed Work……………………………………………………………….. 80
5.2.1 Optimal Healthcare Classification Multi-Disease Diagnostic Architecture (OHCMDD)……………………………………………. 80
5.2.1.1 Data Aggregation Layer………………………………………. 81
5.2.1.2 Data Storage Layer……………………………………………. 81
5.2.1.3 Data Management Layer……………………………………… 81
5.2.1.4 Data Analysis Layer………………………………………….. 81
5.3 Proposed Approach………………………………………………………… 82
5.3.1 D-SMO……………………………………………………………….. 85
5.3.2 Spider Monkey Optimization…………………………………………… 87
5.3.2.1 Foraging Behavior 88
5.4 Experimental Setup……………………………………………………….... 94
5.5. Result and Discussion……………………………………………………… 95
5.5.1 Comparison of the proposed model with Existing Features and Proposed Features……………………………………………………………….. 95
5.5.2 Comparison of Classification Algorithms…………………….....…….. 97
5.5.3 Comparative Analysis with Existing Studies…………………………… 98
5.6 Summary.…………………………………………………………………….. 101

CHAPTER 6: eHealth Architecture for Health 4.0

6.1 Introduction………………………………………………………………… 102
6.2 Features of I4.0……………………………..……………………………… 103
6.3 Supportive Technologies for H4.0 via I4.0.………………..………………….. 105
6.3.1 Adaptive Robotics….…………………………………………………… 105
6.3.2 Embedded Systems…………………………………………………….. 105
6.3.3 Additive Manufacturing……………………………………………….. 106
6.3.4 Cloud Technologies……………………………………………………. 106
6.3.5 Virtualization Technologies……………………………………………. 106
6.3.6 Data Analytics and Artificial Technology……………………………… 107
6.3.7 Communication and Networking………………………………………. 107
6.3.8 Cyber Security………………………………………………………….. 108
6.4 Design Principles of I4.0 to Implement in the Health Sector…………………. 109
6.4.1 Interoperability………………………………………………………….. 109
6.4.2 Virtualization…………………………………………………………… 109
6.4.3 Real-Time Capability…………………………………………………… 109
6.4.4 Decentralization………………………………………………………… 109
6.4.5 Modularity……………………………………………………………… 110
6.4.6 Service Orientation……………………………………………………... 110
6.5 Technology Roadmap to Achieve H4.0 by Implying I4.0…………………… 110
6.6 Smart eHealth Architecture of H4.0 by Implying I4.0………………………… 113

6.6.1 Data Collection Layer………………………………………………… 113
6.6.2 Data Access Layer……………………………………………………. 114
6.6.3 Data Implementation Layer…………………………………………… 115
6.6.4 Data Application Layer……………………………………………….. 115
6.7 Summary………………………………………………………………………. 117
CHAPTER 7: CONCLUSION AND FUTURE WORK
7.1 Summary of Work Done in the Thesis………………………………………... 118
7.2 Contribution of the Research………………………………………………. 119
7.3 Future Work……………………………………………………………………... 120
Publications from the Thesis…………………………………………………… 121
References……………………………………………………………………… 123

List of Abbreviations

ICTs	Information and Communication Technology services
I4.0	Industry 4.0
H4.0	Health 4.0
IoE	Internet of Everything
CPS	Cyber Physical Systems
MIoT	Medical Internet of Things
RAMI 4.0	Reference Architectural Model
ICD-9	International Classification of Diseases
HDFS	Hadoop Distributed File System
HCMP	Healthcare Multi-Phase Architecture
KNN	K Nearest Neighbours
CKD	Chronic Kidney Disease
OHCMDD	Optimal Healthcare Multi-Disease Diagnostic
D-SMO	Density-Based Features Selection with Spider Monkey Optimization
SMO	Spider Monkey Optimization
LR	Logistic Regression
NN	Neural Network
SVM	Support Vector Machine
EHR	Electronic Health Records
BPA	Big Data Predictive Analytics
EMR	Electronic Medical Record
OMDSS	Online Medical Decision Support System
ECG	Electrocardiogram
BDA	Big Data Analytics

WBASN	Wireless Body Area Sensor Network
VPH	Virtual Physiological Human
IoT	Internet of Things
IoS	Internet of Services
PD	Parkinson Disease
BAN	Body Area Network
CoT	Cloud of Things
WHO	World Health Organization
AIDS	Acquired Immune Deficiency Syndrome
NHP	National Health Policy
SBA	Swachh Bharat Abhiyan
SNA	Swasth Nagrik Abhiyan
YARN	Yet Another Resource Negotiator
ESRD	End Stage Renal Disease
RF	Random Forest
NB	Naïve Bayes
GPS	Global positioning system
MRI	Magnetic Resonances Imaging
STL	Standard Triangulate
AM	Additive Manufacturing
VR	Virtual Reality
AR	Augmented Reality
RFID	Radio Frequency Identification
TP	True Positive
FP	False Positive

List of Figures

Figure 1.1: Three Layers of RAMI 4.0 Explains the Way to Handle the Problems of I4.0 in a Structured Manner………………………………………………………… 6

Figure 2.1: Functional Flow Diagram of the Literature Review Phases…………………… 17

Figure 2.2: Bifurcation of Selected Topics for this Study Along with the Count and Then These Topics are Further Divided into Subcategories By Applying IF…………………… 25

Figure 2.3: Keywords Allied to Enabling Technologies…………………………………… 25

Figure 2.4: Keywords Allied to Supporting Technologies of Industry 4.0…………………. 26

Figure 3.1: Representation of Different Sources of Data in The Healthcare System……….. 39

Figure 3.2: Representation of the Working of an Experimental Model…...………………... 43

Figure 3.3: Graph of Calculated Recall as Performance Measures………………………… 44

Figure 3.4: Representation of Precision for Various Models……………………………….. 45

Figure 3.5: Comparison of the Result of the Proposed Study with N.Yuvraj et al. [108] 46

Figure 4.1: The Working Model of HCMP Architecture in Its Six Layers…………………. 52

Figure 4.2: Phase One of Data Distribution ……………….……………………………... 54

Figure 4.3: Working of the Second Phase for Data Distribution …………...……………… 55

Figure 4.4: Working of the Third Phase, when a New DataNode is Admitted in the Existing Cluster 57

Figure 4.5: Algorithm used to Extract the Symptoms of Similar Patients Based on Scores 58

Figure 4.6: Pseudocode of MySymptom Algorithm…………………………………….. 60

Figure 4.7: Execution Time of DataNodes for: (A) The First Round, (B) Second Round Showing the Results for The New Job Type and (C) Third Round Showing Performance of the Hadoop Cluster After Horizontal Scaling…………………. 61

Figure 4.8: Working of Distributed Model for Faster Data Processing……………………... 62

Figure 4.9: Statistical Distribution Representing Kidney Donation in India………………... 65

Figure 4.10 (A): AUC of Decision Tree…………………………………………………….. 69

(B): AUC of Simple Distribution Naïve Bayes………………………………… 69

(C): AUC of Naïve Bayes Kernel………………………………………………. 70

(D): AUC of KNN……………………………………………………………… 70

(E): AUC of Random Forest………………………………………………….... 71

(F): AUC of Random Tree…………………………………………………….. 71

Figure 4.11: Comparison of Performance Measures with Other Models……………………. 76

Figure 5.1: Seven V's of Big Data in Association with the Latest Technology to Achieve H4.0…………………………………………………………………………….. 79

Figure 5.2: Use of Different Technologies to Achieve the Goal of H4.0…………………. 80

Figure 5.3: Optimal Healthcare Classification Multi-Disease Diagnostic Architecture……. 82

Figure 5.4: Flowchart Explaining Data Placement Strategy in Hadoop Cluster By Considering Different Situations……………………………………………… 83

Figure 5.5: Flowchart of Hadoop Based D-SMO Algorithm…………………………….. 87

Figure 5.6: Flowchart Showing the Foraging Behavior of SM……………………………… 88

Figure 5.7: (A) Shows the Important Features of CKD (B) Presents the Important Features of Heart Disease (C) Gives the Vital Features of Diabetes Disease as Selected by DBFS Algorithm………………………………………………………………. 96

Figure 5.8: Comparison of D-SMO, ACO, SMO, DT and RF with respect to (A) Accuracy, (B) Precision (C) Sensitivity (D) F1-Score…………………………………….. 98

Figure 5.9: Performance Comparison of OHCMDD Architecture with [108, 110, 111, 244] (A)Accuracy (B)Precision (C) Sensitivity (D)F-1……………………………… 100

Figure 6.1: All Industrial Revolutions from I1.0 to I4.0 and all health revolutions from H1.0 to H4.0……………………………………………………………………….. 103

Figure 6.2 Features of I4.0 as Vertical Integration for the Self-Organized System, Horizontal Integration for an Efficient Ecosystem Among Various Organizations to Share Information and End-to-End Integration Enables Reusability of the Continuous Product Model………………………………………………………………… 104

Figure 6.3: Different Technologies that Support the Healthcare Industry to Achieve the Goal of H4.0……………………………………………………………………….. 105

Figure 6.4: (A) Image Depicts the Number of Communication Paths Between the Individuals or Doctor, Patient, Staff (B) Communication Channel Increases as the Number of Individuals Increase in this Channel Who Needs to Communicate…………………………………………………………… 108

Figure 6.5: Proposed Technology Roadmap for H4.0……………………………………….. 113

Figure 6.6: The Proposed eHealth Architecture for obtaining H4.0…………………………. 117

List of Tables

Table 2.1: Comparison of Related Literature with the Proposed Work………………. 21

Table 2.2: Classification of Technology Used for I4.0 to Achieve H4.0 and Design Principles……………………………………………………………….. 26

Table 2.3: Summarizes Details of the Selected Articles……………..………………… 27

Table 3.1: List of Big Data Analytics Tools with Their Uses…………………………… 40

Table 3.2: Calculated Recall for Various Models Using Classification Techniques 44

Table 3.3: Calculated Precision for Various ML Models Using Classification Techniques………………………………………………………………. 45

Table 4.1: Representation of Scores Given To Patient Based on Similar Symptoms….. 59

Table 4.2: Explanation of All Rounds Based on Jobs Distribution Performed in DataNodes by NameNodes CR Calculation After Applying Hadoop Distribution Algorithms……………………………………………………. 60

Table 4.3: Parameters Responsible for Chronic Kidney Disease………………………. 63

Table 4.4: Techniques and Methods Used at Different Layers of HCMP Architecture. 65

Table 4.5: Specification of DataNodes and The Experimental Environment of The Hadoop Cluster………………………………………………………… 66

Table 4.6: Comparison of Various Prediction Algorithms for Chronic Kidney Disease. 68

Table 4.7: Confusion Matrix of Different Classifiers………………………………….. 72

Table 4.8: Comparison of Classification Accuracy, Recall, Precision and AUC………. 75

Table 5.1: Configuration of The DataNodes Employed in The Cluster………………... 95

Table 5.2: Presents the Result for CKD, Diabetes and Heart Disease With the Existing Features of the Dataset And The Proposed Features……………………… 97

Table 5.3: Analyzing the Accuracy, Precision, Sensitivity And F1-score of CKD, Diabetes and Heart Disease With Different ML Algorithms……………….. 97

Table 5.4: Analyzing the Performance of OHCMDD Architecture’s Prediction for CKD, Diabetes and Heart Disease With Existing Studies..………………….. 100

Table 6.1: Year-wise Roadmap Techniques Along with Their Explanation…………… 112

Table 6.2: Benefits and Challenges of Roadmap Techniques…………………………. 112

List of Algorithms

Algorithm 4.1: Initial_Data_placement…..…………………………………………….. 55

Algorithm 4.2: Decision_on_Capacity_and_Data_Placemnt……….…………….......... 56

Algorithm 4.3: MySymptom ……………………………….………………………. 60

Algorithm 5.1: Upload_Data_on_HDFS………………..……………..…….......... 84

Algorithm 5.2: DataNode_Insertion_DataNode_Deletion……………………………… 86

Algorithm 5.3: Position_Update_Method_for_LLP……………………………………. 90

Algorithm 5.4: Position_Update_Method_for_GLP……………………………………. 90

Algorithm 5.5: Working_of_LLDP………………………………………………......... 92

Algorithm 5.6: Working_of_GLDP…………………………………………………..... 92

Algorithm 5.7: Working_of_SMO……………………………………………………… 93

CHAPTER 1

Introduction

The chapter begins with introductory details of healthcare big data, and highlights the different axis of big data, Industry 4.0, Health 4.0, RAMI 4.0, and data management. It also discusses the concept and requirement of healthcare, research motivation, problem statement, and contributions. The chapter concludes with the organization of thesis.

1.1 Introduction

The Information and Communication Technology services (ICTs) work as a building block for the healthcare domain [1]. The adoption of ICTs brings new opportunities and discloses unforeseen and novel application scenarios [2, 3]. Due to this, the overall healthcare sector is benefited, as the healthcare services are enhanced with reduced cost [4]. It is possible due to the increased demand for healthcare services for new environmental vulnerabilities and the ageing population. The current healthcare system is designed using advanced competence technologies such as big data analytics, machine learning, and information technology to provide progressive and intelligent healthcare services [5, 6, 7]. The available healthcare services suffer from various issues such as lack of healthcare data handling, data filtering, and right treatment due to which patients are suffering worldwide [8]. A literature survey of the proposed research area underpins the novel and dynamic direction of research fueled by promising technologies and applications. This section presents the details of Health 4.0 (H4.0) along with the 8-P's of healthcare, Reference Architectural Model (RAMI 4.0), different challenges of big data management and Industry 4.0 (I4.0).

1.1.1 Industry 4.0

The industrial revolution appeared as a transformation that brings improvement in manufacturing, healthcare, and service systems [9]. With this improvement, rapid and remarkable changes have appeared in many different sectors such as manufacturing, services, and information technology [10, 11]. Due to these changes and integration with the advancement of ICTs the synergy was aroused among the different sectors [12, 13]. This advancement leads to increase productivity in the manufacturing, service, and healthcare sector. In recent years the service system, manufacturing companies, and healthcare sector have faced considerable challenges due to the coordination of disruptive concepts that appeared after the advancement of the internet in the form of embedded systems, communication and networking, data analytics, adaptive robotics, additive manufacturing, artificial intelligence, etc. [14, 15]. This advancement becomes the main cause of the development in information technology, manufacturing, and healthcare products advance services. The coordination and communication among these technologies constitute the term, I4.0, and H4.0[16, 17, 18].

I4.0 was first announced by the German government as a new industrial revolution and as a key initiative [19]. The I4.0 works in creating smart and communicative systems for human-to-machine interaction and machine-to-machine communication that brings improvement in service, production, and the healthcare sector [20, 21]. In today's scenario and for the future purpose, different organizations have to deal with the management of the flow of data that relied on the assessment and acquisition of data extracted from the interaction of distributed and intelligent systems. The idea of this healthcare data acquisition, management, and processing enables the healthcare sector for parallel data processing to produce the fast result on time. Thus, the healthcare sector has been probing the right version of I4.0 to improve its supply chain, production facility, service systems, and assisted healthcare that enables it to become a value-added healthcare network and achieve the target of health 4.0 [22, 23, 24].

1.1.2 Health 4.0

The H4.0 approach is derived from the renowned I4.0. I4.0 brings revolution in the manufacturing sector with the advancement of ICTs and digitization [25, 26, 27, 28]. The health sector is becoming more personalized and timelier after the advancement of new wearable technology for health monitoring and smart devices [29, 30, 31]. The H4.0 approach focuses on clinical setting, providing assisted healthcare services, and better clinical outcome that stabilizes the healthcare system [32, 33]. The new H4.0 initiative is based on the services and technologies that continue to grow and attract a large number of funds globally and research investment [34, 35].

The main aim of H4.0 is to provide improved healthcare services and allow progressive virtualization that enables the personalization of healthcare, real-time monitoring of patients, and a linked value chain among various organizations [36, 37, 38, 39]. The personalization facility will be achieved for healthcare with the improved ICTs such as cloud computing, Internet of Everything (IOE), Cyber-Physical Systems (CPS), 5G mobile communication network, and services [40, 41]. This is all possible only by applying the design principles of I4.0 to provide improved healthcare facilities [42].

1.1.2.1 H4.0 Towards Improved Healthcare

H4.0 is utilized for leveraging the concept of virtualization and individualization through various industrial domains [43, 44]. The strategies of H4.0 also empower the industries from the manufacturing sector to service providers. It pays more attention to personalization, and individualization services to patients, customers, and clients [45, 46]. This vision of H4.0 helps to provide tremendous progress towards 8-P of healthcare [47, 48, 49].

- **Pervasive and Pre-emptive**: In today's world modern societies are suffering from a sudden increase in fundamental health challenges like lack of medical professionals and proper infrastructure for chronic diseases management [50, 51]. The spread of communicable diseases and non-communicable chronic diseases becomes a serious social problem for the entire world. There is a huge requirement for proper disease management to improve quality of life, and decrease social problems and economic burden. Generally, these diseases are the complex interaction between environmental factors and the genetic make-up of individuals [52, 53]. The pre-emptive healthcare facility takes personal data, physiological information, genetic data, and lifestyle information to predict the complications and risk of illness [54, 55]. The H4.0 can provide a ubiquitous environment for healthcare to monitor and support patients' conditions and helps to avoid expensive hospital care [56]. Pervasive healthcare is utilized to provide preventive and emergency healthcare services using ubiquitous computing technology. IT works towards proactive healthcare facilities that ensure preventive patient care, significant vitals of patients, and continuous monitoring solutions to improve system efficiency and reliability [57, 58].

- **Preventive and Prediction:** Medical Internet of Things (MIoT) is a part of a new ambient healthcare environment that provides (P¬2) health protection through preventive and prediction computing [59, 60]. Personalized health monitoring can be managed with a plethora of health sensors [61]. They help to collect the most critical parameters from an ambient domain that helps to provide P2 health facilities to the patients [62]. The collected data is to be measured in a way to deliver personalized patient care. The interoperability of IoT and medical devices is able to reach more complete and accurate datasets. This facilitates the quick diagnosis of disease and better service. H4.0 with RAMI 4.0 [63] are the key aspects of the system that are determined by prediction and preventive techniques in healthcare and also help to link the service providers to the patients directly [64].
- **Precision and Personalized:** H4.0 with the help of RAMI 4.0 able to provide personalized healthcare facilities with a high value of precision [65]. For this wearable technology plays a vital role as a patient's fitness and health can be monitored and tracked [66]. The health of elderly persons can also be easily maintained using this technology and hospital visits can be reduced [67]. Personalized healthcare enables pre-emption of diseases, early detection, and remote monitoring of disease could be possible [68]. The collected information is directly monitored by the experts and an analytical study helps to provide personalized aids with a high degree of precision.
- **Patient-Oriented and Participatory:** Nowadays mHealth becomes the solution for individuals to manage and monitor their chronic health conditions [69]. The self-tracking applications are utilized in biomedical and observational research where individuals can easily maintain self-care [70]. Through this, they can continually monitor their physiological health and critical biomarkers [71]. This participatory approach provides the point-of-care diagnosis with real-time monitoring for better health conditions [72].

1.1.3 RAMI 4.0

The RAMI 4.0 is a three-dimensional map that simplifies the issues of I4.0 and presents the strategy to handle them in a structured manner as shown in Fig. 1.1. It makes sure that all participants of I4.0 must know everyone's requirements for making things simpler [73]. The

RAMI 4.0 model is a service-oriented model that includes IT security and data privacy. The three-axis of RAMI 4.0 are:

- **Hierarchy Level:** The hierarchical level of RAMI 4.0 presents a functional assignment of components for I4.0. This axis follows the IEC 62264 and IEC 61512 standards within an enterprise [74]. The exceeding top and below limit values of the IEC standards zone resembles the procedure that explains the collaboration among external engineering firms, customers, groups of factories, and component suppliers. Therefore, the hierarchy level performs mapping of interface and standards and executes logical grouping of functions that represents the different functionalities within factories [75]. The model is about: control devices, products, stations, field devices, enterprises, work centres, connected world.
- **Architecture Layers:** The architecture layers include six layers that are business, integration, asset, communication, functional, and information [74]. The layers consistently allow the advancement of I4.0 by interconnecting the mutually dependent manufacturing operations from the point of view of physical and digital world. This representation originates from ICT, where complex properties of a system are broken down into layers.
- **Lifecycle Value Stream:** The left horizontal axis is divided into two parts known as “type” and “instances”. Further, the type is divided into two parts “development” and “maintenance”, on the other hand, an instance is divided into “production” and “maintenance” [76]. A "type" is going to become an "instance" after completing the prototyping, design, and production of the required product. Therefore, "type" is about the initial idea and the manufactured product is the "instance" of a particular type. So, the value stream is represented as the total digitized production in aggregation with purchasing, planning, production, quality, logistic, supplier, and customer.

1.1.4 Big Data

Big Data refers to datasets that are too large and beyond the capacity of classic database software tools to perform basic operations like capture, store, manage, and analyze [77]. It can also be generalized as a vast amount of data, which is nearly terabytes (1012bytes), petabytes (1015bytes), and zettabytes (1021bytes) [78]. This happens due to the advancement of ICTs that help people to produce, share, and process a high volume of data [79]. The different private and public sector industries are also using big data after performing certain operations to improve the services provided by them [80]. The healthcare organization is a multi-dimensional system that also aims to provide better services through fast diagnosis, prevention, and improved treatment of various diseases. Medical data is also present in different formats that are produced from various sources such as patients' medical records, hospital records, medical examination results, health IoT devices, etc. This medical data required proper management and analysis for extracting meaningful information [81]. There are some associated challenges with the handling of big data and they can only be surpassed by using some good data management techniques and BDA techniques [82].

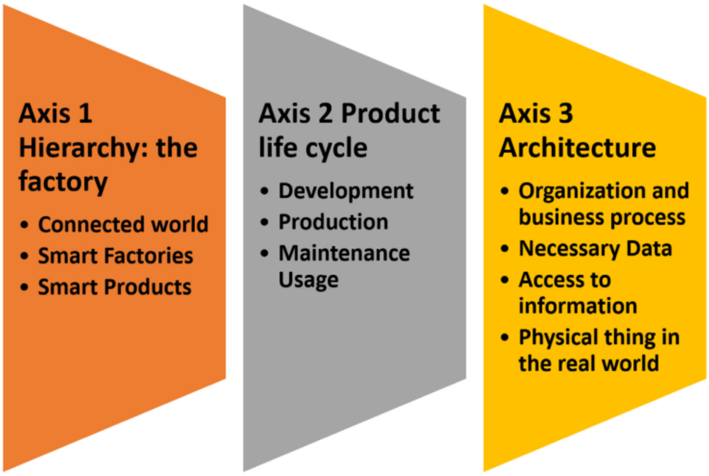


Figure 1.1: Three Layers of RAMI 4.0 Explains the Way to Handle the Problems of I4.0 in a Structured Manner

These are used to improve healthcare services to provide better public health facilities with fully equipped architectures that support healthcare organizations' infrastructure to systematically produce and analyse healthcare big data. The emerging challenges or different axis of healthcare big data are follows:

- **Quantitative Human Phenotyping:** In genetics, the observable characteristics of an organism are known as phenotypes. The phenotype of a person may include information regarding anthropometrics that stores the height and weight measurements, and incessant physiological measurement that contains data collected by wearable devices like a heartbeat, blood pressure, etc [83,84]. Clinical phenotyping includes medical imaging, medication use, and procedure results. In the construction of a patient's information cohort, the choice to keep the depth of information and type depends on different considerations. Providing quantitative human measurements is a challenging task but it helps in delivering a patient-specific environment to prevent the disease and diagnose it in advance [85,86]. The critical thing is to send the available cohort of information to physicians and hospitals for further analysis. H4.0 helps to fill the gap and provides insightful measurement in real-life to provide proactive health services.

- **Communication**: In healthcare, various individuals take part in the interaction to provide a better healthcare facility to patients. This communication may take place between physically distributed systems with the help of improved networking and communication technology to deliver a connected healthcare system [87]. This networking and communication channel works toward longitudinal follow-ups [88]. Longitudinal follow-up contains the information of time intervals or total duration of follow-up of the meeting of patients and doctors or hospitals with manufacturing companies, and pharma companies through linked data sources.
- **Linked Data Sources:** The capability to link various data sources, and then acquire the related information for a person is also a challenging task [89]. For this linkage purpose, the patients should get full ownership of their data to make decisions on data sharing [90]. Then the shared data will be linked to the associated medical information's central repository. Secondly, data privacy should be maintained and linked with the unique Aadhar number of that person. This centralized repository should be attached to the various government schemes like CGHS (central government health scheme) and health insurance data. It will be helpful in the long term when the particular person's data updated as soon as a new development in the information found. This data will be going to utilized by the health practitioner to produce a health report card for that person. The associated healthcare organizations also utilize the information to provide customized healthcare services by considering all parameters of heterogeneity in the population. [91].
- **Heterogeneity of the Population Cohorts:** Heterogeneity in the medical data set refers to various factors including age, gender, race, disability status, ethnicity, educational level, socioeconomic status, and geographical location. The creation and organization of a cohort that represents the real-world population is a challenging task [92]. The inclusion process of the data in a cohort involves various steps like the data of related patients should be included in a common database and the consent of the patient. The data should be refined before storing in the related cohort so it can be properly analysed and should not contain any biased information. In US Research Program they explicitly fix their goal of heterogeneity, with almost 80% of the participants selected from the

historically under-represented groups [93]. These types of cohorts help the organizations to understand the inside pattern of the related group so that person-specific requirements can be fulfilled.

- **Standardization of the Data:** The health dataset may contain different data sources [94]. It is necessary to have the health dataset in a standardized form for further clinical research and comparisons concurrently by researchers around the globe [95]. Thus, toning of data is important and should be performed with a common vocabulary. So, the collection of data must be performed using a standardized ICD-9 diagnoses code and also categorized by standard definitions. This standardized dataset in a universal format allows collaboration across the globe [96].
- **Early Warning in Chronic Disease Monitoring:** It is very important to generate early alarms before the commencement of any chronic disease. It requires active and sustained participation of patients through digitally recording their daily lives in the form of a varying number of details for different purposes. This quantitative self-tracking is the personal behavior of various individuals to track their physical, biological, environmental, and behavioral information [97]. Other measures like quality of sleep, energy level, weight, mood, time use, etc., can also be tracked. The proper management of this information can be utilized by the physician to monitor the health of a person and if something wrong is seen in the data then further testing should be carried out [98]. This flexibility in the healthcare services is only being provided by the H4.0 and implemented successfully with the help of RAMI4.0. Some diseases like diabetes, Chronic Kidney Disease (CKD), heart disease, asthma, etc., require the full participation of patients. Asthma is a prevalent and expensive disease in the United States and it is tough to cure it permanently but using advanced technology the adverse events caused by environmental triggers can be avoided through appropriate drugs [99]. Some researchers like Ram et al. [100] introduced methods to predict visits to the emergency department for asthma in a particular area based on multiple data sources such as social media data and real-time environmental data.

1.1.5 Data Management

Maintaining the healthcare data using conventional approaches has achieved limited success due to their incapability of handling high volume complex data, high variety, and high-velocity data [101]. To overcome this problem of emulated healthcare big data can be managed by BDA platforms like Hadoop Distributed File System (HDFS). HDFS provides a fault-tolerant environment and supports a better replication factor [102]. The HDFS contains NameNode and DataNodes where NameNode works as the master and DataNode work as slave nodes. HDFS provides data integrity and data-parallel processing that handles the volatility of data. Hadoop works on four layers namely; the data storage layer, data processing layer, data access layer, and data management layer. Every layer includes different components of the Hadoop ecosystem [103].

- **Data Storage Layer:** To provide robust data storage and management Hadoop uses HDFS. The main aim is to provide minimum latency for distributed processing [104]. This is achieved by the data locality concept on the DataNodes and executing maps and reducing processing tasks. The Hadoop ecosystem has two data storage components; HDFS and HBASE.
- **Data Processing Layer:** To process the data Hadoop ecosystem uses MapReduce and Yet Another Resource Negotiator (YARN). MapReduce process the data by running a series of jobs known as a map and reducing to perform the queries on the data. Moreover, YARN is also used for data processing and removes the drawbacks of MapReduce by introducing a resource manager and node manager instead of a job tracker and task tracker in MapReduce. Due to this, YARN is more scalable as it can manage a greater number of DataNodes in a Hadoop cluster as compared to MapReduce.
- **Data Access Layer:** To access the terabyte or petabyte of data Hadoop ecosystem has different components; Hive, Avro, Pig, Mahout, Drill, Sqoop, and Flumes.
- **Data Management Layer:** Basically, this layer is a part of modern data management in the Hadoop ecosystem and it is used for managing an enormous amount of data. These are the different components of the Hadoop data management layer; Oozie, Fumes, Ambari, Zookeepers, and HCatalog.

1.2 Motivation & Scope

Big data in healthcare has emerged as a prominent direction of research with promising applications and scientific trials being explored substantially. It came out as a stirring trend with a gamut of demanding applications that range from H4.0, healthcare recommender systems, disease prediction, expert finding, assisted living, etc. The thrust area is to understand

and explore the big data and healthcare framework, its application, and goals toward health 4.0 (H4.0). The requirement is to go through a review and literature survey of significant research of the particular area by expounding its basic terminology and tasks. The idea is to envision "healthcare" by advancement due to the internet, mobile phones, and wearables data, better healthcare facilities could be provided.

The motivation is based on the syndrome from which the healthcare sector is dealing **“Data rich but information poor**". The proposed study provides a method that properly stores and manages healthcare data so that it can be used to predict the disease in advance and this helps to provide improved healthcare services which lead it towards H4.0 by implying the principles of industry 4.0 (I4.0). The initiative is a big step to transforming the country into a digitally empowered healthcare system that ensures the improvement and benefits of every sector of healthcare through the advancement of ICTs.

1.3 Problem Statement

Statement of Research Question

"Can the voluminous healthcare dataset be managed and mined to gain insights for comprehending the paradigm shift from the conventional healthcare system to health 4.0?"

In response to the identified requirement to well exploit the knowledge in the form of information extracted by the disease dataset, this research question can be further divided into the following three sub-questions, and every sub-question will be addressed by this research:

- How does big data help towards H4.0?
- How the architecture helps the disease prediction model?
- How H4.0 be achieved with I4.0?

1.4 Research Objectives

The objective of this thesis is to find techniques to provide a better healthcare system. It specifically aims at developing a healthcare architecture that facilitates the healthcare organization by providing improved storage and analysing methods. These architectures help in achieving the goal of H4.0.

Consequently, the five main research objectives of the work undertaken are:

- To study big data analytics in healthcare.
- To implement and validate various algorithms based on performance measures.
- To develop a healthcare multilevel architecture for storing, managing, and analysing the data.
- To implement a disease prediction model based on the proposed healthcare architecture.
- To compare the proposed work with the existing architecture or algorithm.

1.5 Contribution of Thesis

The objective of this thesis is to find techniques to develop a smart architecture for healthcare that manage the data and perform the analysis task for disease management. The proposed architecture implements Hadoop based data placement strategy for data management and provides an efficient data placement method. Furthermore, the disease data set was analysed by applying machine learning algorithms. The proposed study provides the means to achieve the goal of H4.0 to improve the overall healthcare environment.

The thesis contribution is as follows:

1. The literature review of big data analysis has been performed for healthcare technology to highlight the challenges of the healthcare industry and identify the various mechanism to overcome these challenges.

2. We proposed a novel healthcare multi-phase architecture (HCMP) to predict chronic kidney disease. The HCMP architecture works on six different layers namely: data-collection, data-storage, data-management, data-processing, data-analysis, and report-generation. The data-storage and data-management layers were performed on heterogeneous Hadoop cluster and the profiling methods were used to consider three situations for calculating the capacity ratio of each DataNode in the cluster. MapReduce is used for parallel data processing. Furthermore, horizontal scaling is performed in the Hadoop cluster, and the performance of every DataNode is investigated based on a capacity ratio. The data-analysis layer has performed classification tasks using a Decision tree, K Nearest Neighbours (KNN) classification, Kernel distributed Naïve Bayes, Simple distributed Naïve Bayes (NB), Random Forest (RF), and Random Tree (RT). Among these classifiers, Kernel distributed Naïve Bayes has produced the best results. The experiments are performed using tools such as Hadoop and RapidMiner to evaluate and analyse the efficiency and performance of the proposed architecture.

3. The HCMP architecture also deployed the proposed MySymptom algorithm to filter the Chronic Kidney Disease (CKD) dataset of patients according to their symptoms at the data processing layer. The case study of CKD in the Indian environment has been explored.

4. The proposed work is enhanced to analyse multiple diseases using a new Hadoop-based Optimal Healthcare Classification Multi-Disease Diagnostic (OHCMDD) architecture for handling multiple diseases database with reduced features set. The OHCMDD architecture also handles the DataNode deletion problem in Hadoop's cluster. Also, an intelligent classification prediction model is introduced for healthcare, namely Density-Based Features Selection with Spider Monkey Optimization (D-SMO) is used for CKD, heart and diabetes disease. The presented intelligent multi-disease model was implemented using the Hadoop cluster, HDFS, and R platform.

5. The technological developments in healthcare, big data, and I4.0 are individually helping to achieve the goal of H4.0. The proposed work provides the road map for H4.0 along with the different technologies involved in it. The work also provides an e-health architecture that is deployed to achieve the target of H4.

6. The analysis and performance evaluation of the experimental results demonstrate that the proposed work provides a reliable architecture for a better healthcare environment. Moreover, the comparative analysis displays that the proposed work showed an improvement. Thus, the proposed study provides an optimal and effective architecture for the healthcare industry toward H4.0.

1.6 Thesis Organization

The organisation of the thesis is presented in this section which comprises of seven chapters.

Chapter 1: Introduction

This chapter explains the introduction and fundamental concepts of the research area followed by the problem statement. Further, this chapter encompasses the outline of the thesis with a summary of the chapter at the end.

Chapter 2: Related Work

This chapter comprises research methodology by explaining and defining the problem of the research in detail and also formulating research questions. The chapter explicates a fleet study of existing work related to healthcare data management technology and healthcare data analytics to achieve the goal of H4.0. Thereafter, the research gaps are identified and listed based on existing studies.

Chapter 3: Exploring Big Data for Healthcare

This chapter presents the prominence of big data analytics for healthcare. It covers the detailed problems faced in handling healthcare inconsistent data. It explains the big data analytics of unstructured data using different tools. The chapter also describes the healthcare ecosystem followed by the results performed after extensive experiments and assessment.

The following paper has been published from this work

- Sisodia, Amrita, and Rajni Jindal. "Exploring the application of big data analysis in healthcare sector." In 2017 International Conference on Computational Science and Computational Intelligence (CSCI), pp. 1455-1458. IEEE, 2017.

- Sisodia, Amrita, and Rajni Jindal. "Prediction of Environmental Diseases Using Machine Learning." In International Conference on Innovative Computing and Communications, pp. 521-531. Springer, Singapore, 2022.

Chapter 4: Healthcare Multi-Phase Architecture for Disease Prediction

This chapter illustrates HCMP architecture with two new algorithms for handling healthcare data management-related problems followed by the one new MySymptom algorithm for filtering out the data of CKD patients. It presents the complete case study of the CKD problem in India with substantial experiments to predict CKD in patients using ML algorithms. It also presents a comparison with different studies.

The following paper has been published from this work:

Sisodia, Amrita, and Rajni Jindal. "An effective model for healthcare to process chronic kidney disease using big data processing." Journal of Ambient Intelligence and Humanized Computing (2022): 1-17. **(SCIE, IF: 7.104)**

Chapter 5: D-SMO Based Multi Disease Prediction Model

This chapter elucidates a multi-disease diagnosis architecture for healthcare. It presents the details of H4.0 and how this technology handles the challenges of healthcare big data. One new D-SMO algorithm has been introduced along with the data management techniques. The comparison of results with other studies demonstrates the effectiveness of the proposed architecture.

The following paper has been communicated from this work:

“Investigation of multi-disease diagnostic system for healthcare using hybrid D-SMO algorithm”

Chapter 6: eHealth Architecture for Health 4.0

This chapter presents the proposed eHealth architecture to achieve the target of health 4.0 and explains the proposed roadmap for health 4.0.

The following paper has been published from this work

Sisodia, Amrita, and Rajni Jindal. "A meta-analysis of industry 4.0 design principles applied in the health sector." *Engineering Applications of Artificial Intelligence* 104 (2021): 104377. **(SCIE, IF**: **6.212)**

Chapter 7: Conclusion

The last chapter presents the conclusion in conjunction with the future scope of the present research work. It deliberates the importance of the proposed healthcare architecture and data management method for achieving the target of health 4.0.

List of Publications: This section lists published/accepted/communicated papers relating to this research work in International/National Journals/Conferences of repute.

References: This section lists references cited in this research work.

CHAPTER 2

Literature Review

This chapter offers a literature review on traditional healthcare data monitoring and classification of various diseases that motivates the development of medical diagnosis and health monitoring towards health 4.0. It formulates the research questions, identifies & classifies the research gaps, and states future guidelines.

2.1 Introduction

Information technology advancement motivates widespread development in health monitoring and medical diagnosis. The use of IoT technologies, wireless sensing technology, and mobile phones are generating healthcare-related data and the proper maintenance of this data is a big challenge [105]. This literature review aims the study of existing research work, identify the research gaps and propose the appropriate possible solution for them. As shown in Fig. 2.1, the literature review comprises of four phases. The first phase is about the planning of the review process and identifying the various categories of the review. The second phase was performed to formulate the research questions based on healthcare data monitoring and classification techniques for different diseases. In the third phase, the categories of the review were reported and research gaps are identified based on the study. The final phase concludes the review with future directions.

2.2 Planning the Review

This section presents a review of the related work of data distribution and management techniques and different approaches used for the analysis of healthcare data. We have summarized and listed key studies and evaluated these studies for searching the research gaps that exist in the studies. The identified research gaps of the key studies are listed below in section 2.4.3. The key studies are divided into two parts as given below:

- The first part covers the work related to data management along with healthcare data analytics-related problems.
- Second part discusses the use of Industry 4.0 (I4.0) in Health 4.0 (H4.0)

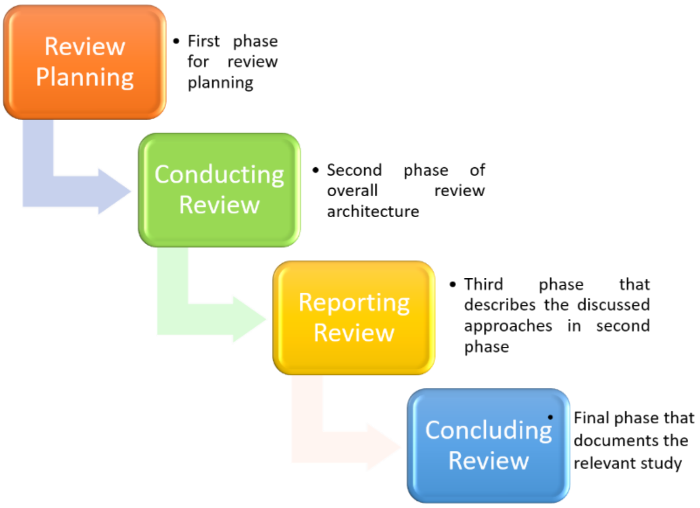

Review Planning
• First phase for review planning
Conducting Review
• Second phase of overall review architecture
Reporting Review
• Third phase that describes the discussed approaches in second phase
Concluding Review
• Final phase that documents the relevant study

Figure 2.1: Functional Flow Diagram of The Literature Review Phases

2.3 Conducting the Review

The purpose of this review process is to develop research questions. The main objective is to present the ongoing research on healthcare data management, healthcare data analytics and H4.0 techniques. Therefore, we framed five research questions as follows:

- **RQ1: Name the various issues from which the existing healthcare services suffer?**

  To identify the healthcare data handling, and data filtering issues for providing the right treatment.

- **RQ2: What are the various strategies required for providing better healthcare data handling issues?**

  To identify various techniques used for better healthcare data handling issues.

- **RQ3 What are the latest healthcare data monitoring and classification methods?**

  To study and understand healthcare data monitoring technology, we collect research papers and mapped the present research area to clarify the healthcare classification and data handling related topics and concepts.

- **RQ4 What are the technological roadmap to achieve health 4.0 by implying industry 4.0?**

  Investigate the technology roadmap to achieve health 4.0.

- **RQ5 What should be the smart eHealth architecture of health 4.0 by implying industry 4.0?**

  Presenting the architecture to achieve health 4.0 via industry 4.0.

2.4 Reporting the Review

In this section, we have described the review of three approaches, healthcare data management techniques along with healthcare data analysis related problems and H4.0 studies. Finally, we have identified the research gaps based on these studies.

2.4.1 Traditional Healthcare Data Handling Techniques and Classification Techniques for Healthcare

In 2020, Seethalakshmi et al. [106] established a novel scheme for big data management. Authors have used a hybrid gradient descent algorithm with spider monkey optimization (SMO) for big data resource scheduling and it also handles the issues faced in heterogeneous Hadoop environment. They have used the foraging behaviour of the SMO algorithm for resource allocation. The algorithm is designed for an efficient task scheduling approach to balancing a load of virtual machines according to their requirement. Also, in 2020, Selvi et al. [107] developed a method using the MapReduce-based classifier technique for the diagnosis of DM disease. The author used different stages of the Hadoop ecosystem using a GBT classifier with an improved K-means clustering approach. Similarly, in 2019, Yuvaraj et al. [108] developed a Hadoop cluster-based data distribution framework for computing DM disease. In 2020, Ramani et al. [109] used a MapReduce-based framework to modify ANN for the prediction of DM. They also use the min-max normalization method for data pre-processing.

To improve healthcare services some researchers worked in the selection process of virtual machines so that healthcare requests can be directly handled. In 2018, Abdelaziz et al. [110], author narrated the challenges faced in the optimal selection of a virtual machine for processing medical requests. They presented that the proper selection of virtual machines is very important as it reduces the execution time to process a medical request. The author uses Healthcare Service (HCS) based model for the cloud environment using PPSO to optimize the virtual machine selection. They have also created a hybrid model for CKD diagnosis using logistic regression (LR) and neural network (NN), where LR is used for feature extraction and NN is used for prediction. CKD is a critical health problem that affects a big range of society and its pervasiveness is constantly increasing. It is very hard to detect CKD without urine and blood test. In 2019, Elhoseny et al. [111] introduced a hybrid scheme for CKD disease. The author tells the importance of the features extraction process for CKD and how it helps in diagnosis by Nephrology staff. For this, they design a density-based feature selection algorithm to extract the best features, and ACO is used for classification. However, in 2017 Polat et al. [112] also used wrapper and filter approaches for selecting the best features of CKD with a Support Vector Machine (SVM). Their result shows that the best accuracy was achieved for SVM with a filtered subset evaluator for CKD prediction.

In 2021, Harimoorthy and Thangavelu [113] provided a general architecture for the multi-disease prediction model. The model was experimented with using CKD, heart disease, and diabetes datasets. An SVM-Radial bias kernel method was used and compared with other ML techniques. In 2019 Makino et al. [114], used Artificial Intelligence (AI) for the prediction of Diabetic Kidney Disease (DKD). The predictive model was created using the Electronic Medical Records (EMR) of 64,059 diabetic patients. AI was used to extract the structural features from the EMR after considering the history. In the study proposed by Wang et al. 2019 [115], a surveillance system for CKD in China was created. Authors have used longitudinal individual EMR data for identifying people suffering from CKD using diagnostic codes and laboratory tests. They developed a clinical decision support system based on real-world data, domain expert knowledge, and clinical guidelines. An online medical decision support system (OMDSS) was developed by Arulanthu and Perumal in 2020 [116] to predict CKD. The model included different stages like the gathering of data, data preprocessing, and data classification. Logistic Regression (LR) was used for classification along with two-parameter twining methods. Adaptive learning rate optimization and adaptive moment estimation algorithm are used to improve the performance of the model. Moreover, a cloud-based IoT multi-layer architecture for medical monitoring is proposed by Asghari et al. in 2019 [117]. The work provides privacy for critical medical data by protecting unauthorized access. The disease prediction model was created using data mining and classification methods. The authors also addressed the side effects and causes of a particular disease. They performed experiments on the Weka mining tool.

In another similar study by Vasquez-Morales et al. in 2019 [118], the authors used a neural network (NN) and a case-based reasoning method (CBR) for prediction. The method was trained using medical care information and patient demographic data. The dataset belongs to two population groups, where one group is diagnosed with CKD and another group without CKD. Their model achieves 95% accuracy with the test dataset and they identified 3,494,516 people at risk of emerging CKD, which is equivalent to the total 7% of the population of Columbia.

At the same time, many researchers use the cross-validation method in their studies to prevent overfitting. In the study by Kriplani et al. in 2019 [119], the authors used a deep neural network to predict the presence of CKD along with a cross-validation technique. They compared their results with other available algorithms. At the same time ensemble approaches also become very popular, as it produces good results in comparison to other ML techniques. In the study by Jongbo et al. in 2020 [120], the authors worked on two ensemble approaches. The study included different algorithms like KNN, NB, DT, Random Subspace Ensemble (RSE) classifier, and bagging ensemble classifier. Among them, they got 100% accuracy with the KNN classifier on RSE. In 2017 a study by Boukenze et al. [121], authors performed their experiments on Weka. They used five different ML models to predict CKD. In the comparative analysis of ML models, they found that Multilayer Perceptron (MLP) and DT (C4.5) produced the best result. Moreover, Sossi et al. in 2019 [122], utilized the IBM SPSS tool to perform their experiments on CKD prediction using ten different ML algorithms. Among all of them, XGBoost Linear scores high accuracy. Table 2.1 represents the assessment of all the articles along with the comparison of the proposed study by considering different points of comparison like different architectures, heterogeneity, energy efficiency, time along with health service, and data handling composition.

Table 2.1: Comparison of Related Literature with the Proposed Work

Reference	**Architecture**	**Distributed Systems**	**Heterogeneous**	**Scalability**	**Learning-based**	**Platform**	**Time**	**Health service**	**Data handling composition**
Seethalakshmi *et al*, 2020. [106]	✗	✓	✓	✓	✓	Hadoop based	✗	✗	✓
Selvi *et al*., 2020 [107]	✗	✓	✓	✗	✓	Hadoop based	✗	✓	✗
Yuvaraj *et al*., 2019 [108]	✓	✓	✓	✗	✗	Hadoop based	✗	✓	✗
Ramani *et al*., 2020 [109]	✓	✓	✓	✗	✓	MapReduce-based	✗	✓	✗
Ahmed Abdelaziz *et al*. 2018 [110]	✓	✓	✓	✓	✓	Cloud based IoT	✓	✓	✗
Elhoseny *et al.* (2019) [111]	✗	✗	✗	✗	✓	MATLAB	✗	✓	✗
Polat *et al.* 2019 [112]	✗	✗	✗	✗	✓	WEKA (version 3.6.13)	✗	✓	✗
Harimoorthy and Thangavelu 2021 [113]	✓	✓	✓	✓	✓	R studio	✗	✓	✗
Makino *et al.* 2019 [114]	✗	✗	✗	✗	✓	Phython	✓	✓	✗
Jinwei Wang *et al.* 2019 [115]	✓	✗	✓	✓	✓	Big data	✓	✓	✗
Arulanthu and Pramila 2020 [116]	✓	✓	✓	✗	✓	IoT and Cloud	✓	✓	✗
Parvaneh Asghari *et al.* 2019 [117]	✓	✓	✓	✓	✓	IoT and Cloud	✓	✓	✗
GABRIEL R. VÁSQUEZ-MORALES *et al.* 2019 [118]	✗	✗	✗	✓	✓	Big data	✓	✓	✗
Himanshu Kriplani *et al.* 2019 [119]	✗	✗	✗	✓	✓	Not mentioned	✗	✓	✗
Olayinka Ayodele Jongbo *et al.* 2020 [120]	✓	✓	✓	✓	✓	Python	✗	✓	✗
Basma Boukenze *et al.* 2016 [121]	✗	✗	✗	✗	✗	Not mentioned	✗	✗	✗
Safae Sossi Alaoui *et al.* (2018) [122]	✗	✗	✗	✗	✓	Not mentioned	✗	✗	✗
Proposed work	✓	✓	✓	✓	✓	Hadoop based	✓	✓	✓

2.4.2 H4.0 Approach for Healthcare

As H4.0. work towards the improvement of healthcare sector by advancing the medical equipment manufacturing and production of medicines [123]. This advancement helps in providing better reliability, quality, process precision, better productivity and cost-effectiveness through the adaptation of robotics technology and advanced automation [124]. This technological innovation in healthcare with robotics helps in reshaping its operational structure [125]. The emerging robotics technology is gaining attention in both academia and industry [126]. In healthcare, robots are used to assist persons with disability, surgeons, medical equipment production industries, etc. For healthcare technology, there is a big range of robots that helps in various places like catheter robotics, surgical robotics, companion robotics, and hygiene robotics [127]. Before performing the procedure, preclinical work was performed in the innovative navigation and imaging laboratory [128].

The involvement of robots in surgeries is becoming more popular due to the preciseness of robots [129]. Robots are leveraged with flexible wristed instruments and 3-D high-definition magnified vision systems [130]. These properties help them to bend, rotate and view clearly in comparison with humans. The usage of surgical robots assists surgeons in conducting invasive surgeries [131].

There is another range of medical-assisted robots used for catheter procedures. In these procedures, the catheter robots perform inside the blood vessels of patients to avoid open surgery [132]. These surgeries are delicate and time-consuming so the surgeons required effective and flexible catheters. In addition, assistive robotics are also used to perform tasks and sensory information that assists senior people and the people with disabilities [133]. The assistive robots enhanced the strength of paralyzed people and in spinal cord injuries of patients to regain the ability to walk [134].

In some cases, the implantable medical devices and wearable devices contain advanced biosignal acquisition, system integration, processing circuits, neural recording, etc. [135, 136]. However, it is quite difficult to overcome from some of the challenges like large health data collection, multi-sensorial data fusion, multi-parameter measurement, energy scavenging, authentication of data, and proper analysis [137]. The automation technologies are also utilized in the production of bio microfluidics droplets used in regenerative medicines, biosensors and drug delivery [138].

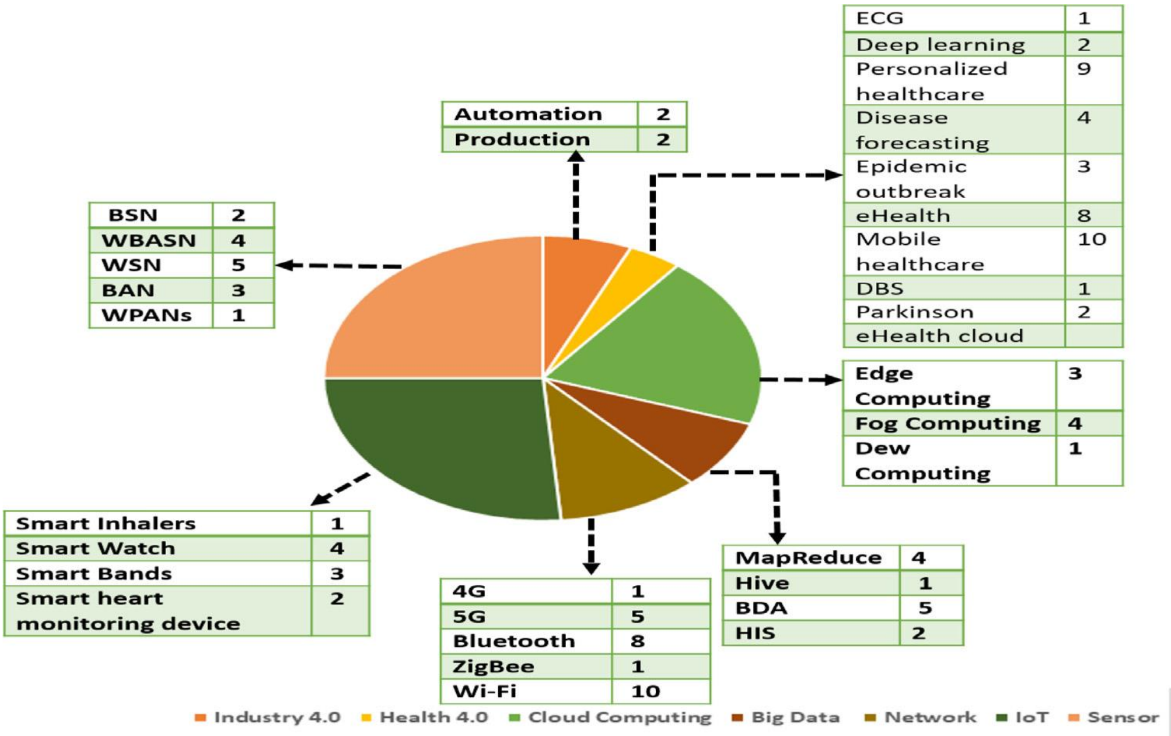
H4.0 also have an impact over automated medical production as a part of advanced manufacturing [139]. This involves the production of medical equipment and medicines. The robotics technology and advanced automation helps in increasing productivity, cost-effectiveness, quality, resource efficiency, reliability and process precision [140]. Healthcare big data is also playing an important role in improving healthcare services [141]. The tremendous volume of healthcare data which relates to patients is gathered from wearable devices [142]. It allows to collect the high volume of patient details on every day basis. The details include disease information, their progress, its causes and cures. This type of structured and unstructured data is collected from different sources which is important in diagnosing patient disease, providing treatment for a particular disease, and projecting the actual condition of a patient's health. This category of diagnosis is also important for providing evidence-based medicine and the production of patients-based implants or healthcare devices [143].

Nowadays for healthcare big data cloud-based information systems are also largely adopted to simplify and strengthen healthcare organizations by helping them in collecting, processing, and deploying clinical records [144, 145, 146]. The cloud-based healthcare architectures extend the data collection process by involving individuals to enter their information through mobile phones and making the information sharing task easy across various involved organizations [147, 148, 149]. Although, data integration and data management are big tasks as different formats of data are collected for increasing the system performance [150].

In extension to this, in a study by Silva et al. [151] author discusses the importance of mobile health for healthcare services regardless of mobility constraints such as organizational barriers, temporal, and geographical. It provides easy access to health services irrespective of place and time and also helps in reducing the high cost of national health services. Instead of this, it empowers the patients and families for self-care in lifestyle-related and chronic disease monitoring by providing better connectivity through the internet [152]. The advent of 4G health defines as the long-term growth of m-health for providing personalized medical facilities [153]. The 5G network is the advanced version of 4G where it integrates wireless access technologies, Wi-Fi, 4G, and m-health paradigm that will be leveraged by the advanced capabilities of 5G [154]. Based on these studies mobile technologies such as mobile apps and smartphones are powerful and act as emerging tools for transferring health information. This technological advancement is generally considered as a pillar for health-information delivery, direct care, m-health, monitoring of physiological signs, patient education, etc. [155].

There is a new amendment in the ICT pillar as ubiquitous health that supports the provision of pervasive healthcare services to anyone, anytime, and anywhere. It ignores the time, location and other factors and increases the quality and coverage of healthcare [156]. Pervasive healthcare focuses on deinstitutionalised healthcare services for patients by providing anywhere and anytime health facilities to reduce the institutionalization cost faced by the healthcare organization [157]. These definitions of pervasive healthcare highlights the social impact of healthcare availability for anyone instead of technological aspects. Whereas in a study by Sobrino et al. [158] they define pervasive healthcare as one that uses the pervasive computing paradigm like IoT, IoE for providing the medical facility at home. These facilities are required for improving medical services and providing sustainable healthy life. For this, a dynamic interconnected network is required that supports healthcare services independent of location and time [159]. Ubiquitous health is considering all the factors of providing better prevention, treatment, diagnosis, and follow-ups facilities without going physically to doctors. The development in the healthcare sector improves the prevention and pre-diagnosis of diseases [160].

These studies demonstrate the use of I4.0 in H4.0 by considering various factors of digitization. The bifurcation of the topics is based on the broad categories related to H4.0, I4.0, cloud, network, sensor, big data, and IoT as shown in Fig. 2.2. The different healthcare ICT-based paradigms are personalized health, mobile health, e-health, ubiquitous health, smart health, pervasive health. All of them are used for the advancement of healthcare organizations. With the help of numerous technologies that are applied to support the understanding of H4.0 via I4.0 are depicted in Fig 2.3. Communication technology, BDA, machine learning as data analytics technology, cloud computing, and network technology are frequently used as keywords in different studies. At the same point keywords like 3D printing, computing and virtualization attain less attention in recent studies. Figure 2.4 provides information about the supporting technology used for health 4.0. It was observed that system integration, human-machine collaboration, industrial automation, safety, and security attain more attention from researchers and in contrast decentralization, customization, and agility get inadequate attention. In this regard Table 2.2 explains the correlation between existing technology and designing principles of I4.0. The considered observation from different studies shows the coordination between supportive technologies and design principles.



Automation	2
Production	2

ECG	1
Deep learning	2
Personalized healthcare	9
Disease forecasting	4
Epidemic outbreak	3
eHealth	8
Mobile healthcare	10
DBS	1
Parkinson	2
eHealth cloud	

BSN	2
WBASN	4
WSN	5
BAN	3
WPANs	1

Edge Computing	3
Fog Computing	4
Dew Computing	1

Smart Inhalers	1
Smart Watch	4
Smart Bands	3
Smart heart monitoring device	2

MapReduce	4
Hive	1
BDA	5
HIS	2

4G	1
5G	5
Bluetooth	8
ZigBee	1
Wi-Fi	10

Figure 2.2: Bifurcation of Selected Topics for This Study Along with the Count and Then These Topics are Further Divided into Subcategories by Applying IF

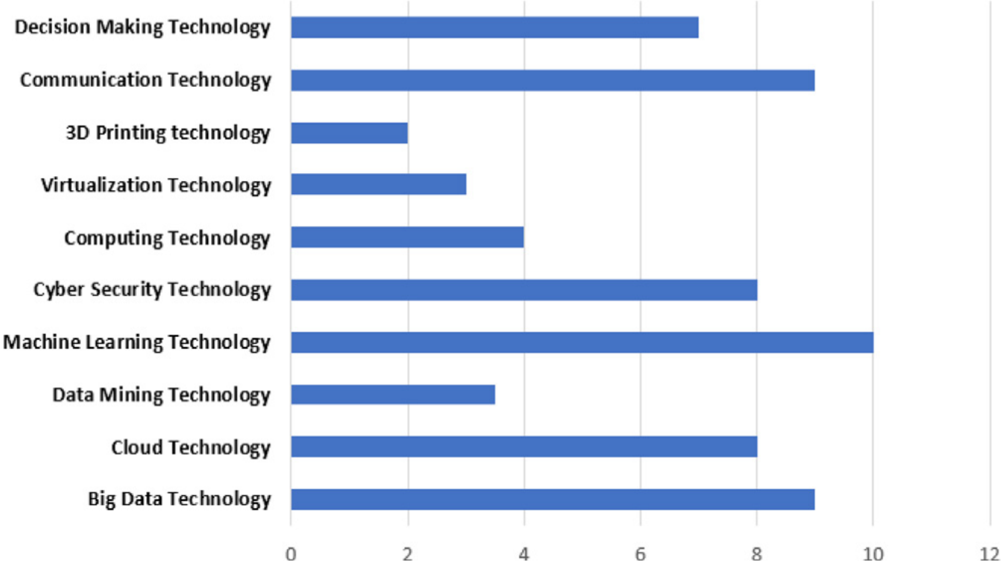


Figure 2.3: Keywords Allied to Enabling Technologies

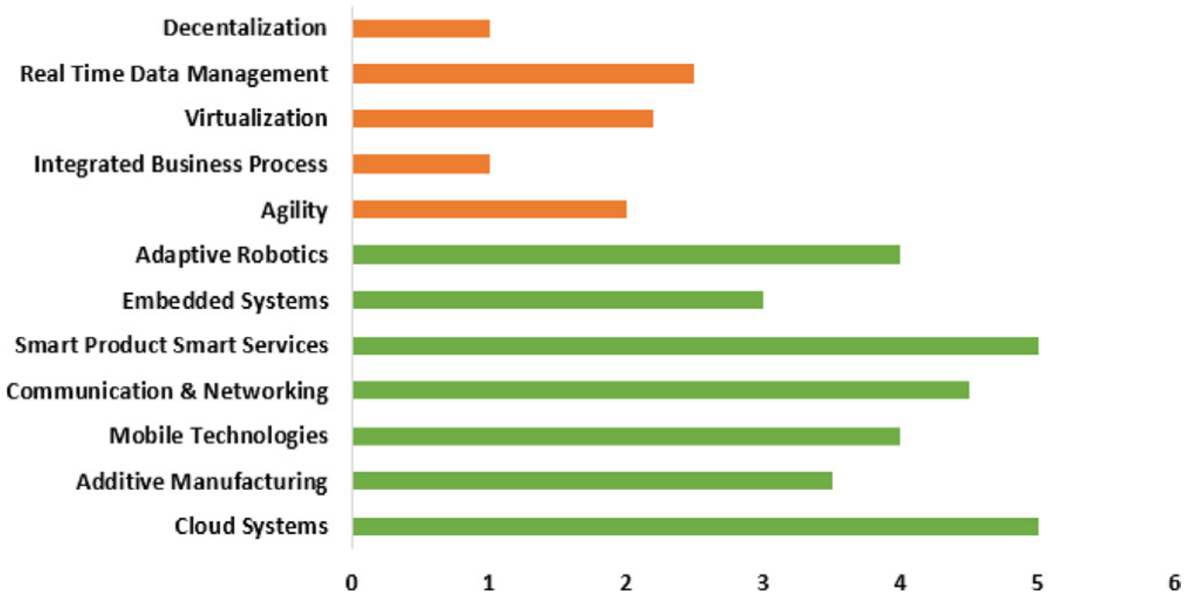


Figure 2.4: Keywords Allied to Supporting Technologies of Industry 4.0

Table 2.2**:** Classification of Technology Used for I4.0 To Achieve H4.0 And Design Principles

Design Principles	**Adaptive Robotics**	**Cyber Security**	**AI & Data Analytics**	**Embedded Systems**	**Communication & networking**	**Cloud technology**	**RFID & RTLS**	**Sensors & Actuators**	**VR &AR**
Interoperability		×			×				
Virtualization		×							×
Decentralization			×	×	×		×	×	
Real-time Capability			×				×	×	
Service orientation			×			×			
Modularity		×			×	×	×	×	

Therefore, the extracted information of the studies includes author details, techniques applied, objectives of the studies, advantages and disadvantages of the selected studies, and finally the device used by them to implement it is presented in Table 2.3. From the selected studies, it was observed that H4.0 can be seen as an integral part of I4.0 where everything is connected to the internet in this digitized world. It is an era of smart factories and the Internet of Everything (IoE) in which people, processes, data, and things are connected. Study shows that due to advancement the healthcare industry is facing lot of changes. Nowadays the basic functionality of the health industry has become more robust so that timely health service is provided to the patients [161]. The second thing is providing service to the patients who are present in remote areas or out hospital services and the last one is the usage of smart devices in healthcare. Robotics help and precision medicine [162] also plays an important role in the advancement of the healthcare industry.

Table 2.3: Summarizes Details of The Selected Articles

Author	Techniques	Objectives	Advantages	Disadvantages	Device Used
Abdelghani Benharref, et al. [163]	Cloud technology.	The cloud-based novel healthcare system is presented for monitoring, tracing, and preventing chronic diseases.	Cloud base system is useful to cater to the storage of large records of patients for processing and managing. This provides faster and more cost-effective services to larger organizations.	The framework is not fully utilized.	Sensor, mobile phone, Wi-Fi.
Aparna Kumari, et al.[164]	Fog, Cloud, IoT.	A three-layer patient-centric architecture was built for real-time data collection, processing, and transmission.	Fog-based systems help doctors to take fast decisions at the time of emergency.	No implementation is done.	Smartwatch, ECG monitors.
Gunasekaran Manogaran et al.[165]	Sensor, IoT, Cloud Computing.	Proposed architecture for secure internet of things for storing and processing large and scalable sensor data.	The architecture is used to send an alert message to the patient if the particular range exceeds, for example, heart rate and blood pressure if exceeds the threshold limit.	Security issues on the cloud are not handled properly.	Wearable sensors.
David W. Bates et al.[166]	Big Data Analytics (BDA).	To use electronic health records for reducing the treatment time and take benefit of BDA.	Big data healthcare is used to reduce the cost of critically ill patients.	No framework is given.	High-power computers.
Christoph Thuemmler et al.[167]	Big Data Analytics (BDA).	By using health 4.0 open ups the gate for smart pharmaceuticals.	To provide customised service to asthmatic patients by providing smart inhalers.	Lots of connectivity problems in practice.	Mobile phone, smart inhalers, Bluetooth.

Sergio Gonzalez-Valenzuela *et al.*[168]	Wireless body area sensor network (WBASN).	Improving wireless communication by decreasing the packet loss rate after introducing a wireless communication protocol.	The wireless communication rate increased.	The energy harvesting schema is not used.	Wireless body sensors, Wearable sensors.
Marco Viceconti, *et al.*[169]	BDA	A method is proposed in the paper to combine big data analytics with Virtual physiological human technologies for producing a robust solution for silico medicine.	An efficient method is defined to improve silicon medicine technology.	No framework is given to handle the problem.	Virtual physiological human technology.
Gunasekaran Manogaran *et al.*[170]	Wearable health sensors.	A scalable body sensor architecture is developed for cloud computing for storing and processing healthcare data.	Big data technologies are used to produce results with better accuracy.	To validate the model only a single parameter i.e. accuracy is used.	Sensors, IoT devices.
Nitesh V. Chawla *et al.* [171]	Personalized healthcare, data mining.	A personalized patient-centred healthcare framework is developed with big data.	Patients' data processing time is reduced, helpful to physicians.	Not useful in every condition.	High-power computers.
Harishchandra Dubey *et al.* [172]	Speech therapies in Parkinson's disease (PD).	A smartwatch-based system is designed to monitor the speech therapy of Parkinson's patients.	Help to assist the patients in the absence of trainers.	Very limited features are available for training purposes.	Eco Wear watch Audible microphone, Smart Phones.
Orestis Akrivopoulos, *et al.*[173]	IoT, Wearable devices, 5G, Fog Computing.	Implementation of a framework by using a middle layer of fog between the end-user and remote cloud.	Fog alleviates the architecture and help to consider various issues like security, bandwidth consumption, and scalability.	Not productive if the number of devices increases.	ECG, wearable devices, Mobile phones.
Teemu H. Laine, *et al.*[174]	Body Area Network (BAN), 4G, 3G, ZigBee, Bluetooth, Mobile phone.	A Bluetooth- gateway system is used to bridge the gap between the internet and the sensor network.	By using the system patients can also update their records and keep watch on them.	The automatic alarm feature is not present for emergency situations.	ZigBee, USB, Bluetooth, BAN.

Abderrahim Bourouisa *et al.* [175]	Smartphone, Tablet, Intelligent system.	An innovative approach is developed to support a low-cost smartphone-based intelligent system that works by using an integrated camera.	Remote area patients can scan their skin and get a treatment related to the disease.	The approach is not applicable to all skin diseases.	Smartphone, Tablet.
P. Kumar *et al.*[176]	BDA, Cloud, IoT.	Enhancing the performance of the healthcare system by reducing execution time, and optimizing the storage of patients' data with a real-time retrieval mechanism.	The new model was built to optimize the selection of virtual machines in IoT and cloud environments for handling and managing a big amount of data.	Cloud and IoT security problems were not handled.	Mobile Phone, Personal Computer, Sensors.
AHMED M. ELMISERY et al. [177]	IoT, Cloud Computing, IBE-Lite Scheme (lightweight identity-based cryptography).	The holistic privacy framework was developed and Implemented as a middleware for collaborative privacy.	Enhanced privacy features help patients in various ways.	A real-time test on datasets is missing.	Sensor nodes, mobile phones.
Harishchandra Dubey *et al.*[178]	Big data, Body area network, Cyber-physical systems, edge computing, fog computing, IoT.	A service-oriented architecture for fog computing is built for telehealth applications.	Overall system efficiency is improved after using fog architecture.	A real-time test on smart homes is missing.	Sensors, IoT devices, mobile phones.
Cristiano André da Costa *et al.* [179]	Bluetooth, WiFi, Smart Health Objects (SHO).	Concept of Internet of Health Thing (IoHT) was introduced and focused on common heuristic approaches.	The study tells that IoHT paradigm provide optimal solutions for managing patients' information in hospital wards.	No security parameter is defined to secure this IoHT data.	Electronic Health Record(HER), Personal Health Record(PHR), IoT, IoHT.
Ana Carolina Borges Monteiro, *et al.*[180]	Hospital Information System (HIS), MIoT, health 4.0, Cloud Computing.	Delivery of the concept of health 4.0 by using 5 G-driven personal care and Medical Internet of Things(MIoT).	A detailed explanation of health 4.0 with the correlation of various supporting technologies.	No means is mentioned to handle real-time health data.	IoT, Mobile phones, Sensors, Mobile Edge Cloud (MEC).
Iuliana Chiuchisan, *et al.*[181]	BAN, IoT, Wireless Personal Area Network(WPANs).	Proposed a solution for monitoring people at risk at preventing them by providing medical help on time.	The use of IoT in e-Health is presented with the help of a case study.	The review is very complex to understand.	IoT, Mobile Phones, sensors.

Mattia Arlotti *et al.* [182]	Conventional Deep Brain Simulation(cDBS), Deep Brain Stimulation(DBS).	Clinical translation of a DBS is presented with its limitations, benefits and unsolved issues.	The study provides some of the very critical issues related to a DBS in a simple manner.	In this review, a very short number of studies are considered related to the topic.	Surface EMG, Embedded device, Wearable device.
Hans Löhr *et al.* [183]	eHealth Cloud	The study presents e-health from the client security platform point of view.	The study presents client platform security in a very good manner.	To validate the model only a single parameter i.e accuracy is used.	EHRs, VM, Trusted Virtual Domains(TVDs).
Amir M. Rahmani *et al.* [184]	Cloud, Fog, IoT.	The concepts of Fog computing are exploited by using healthcare IoT systems to form an intelligent intermediator layer of distribution between sensor nodes and the cloud.	A fog-assisted system is used to develop an intermediator layer.	Data security issues are not handled properly.	Body Sensors, Wi-Fi node, Bluetooth node, BLE node.
George Suciu *et al.* [185]	Cloud Computing, IoT, M2M, Remote Telemetry Units(RMTs).	e-Health architecture was proposed that built on Exalead cloud view environment.	Architecture is present that handles sensor data for supporting RTUs.	The process is very complicated.	Sensor, Wearable, VM.
Farshad Firouzi *et al.*[186]	Big data, Edge computing, IoT, Machine learning.	e-Health and m-Health services are used with IoT to provide personalized and precise care facilities.	The process can be used to provide better health services.	The complex process to understand and implement.	IoT, sensors, Mobile phones, and Tablets.
M. A. de Jesus *et al.*[187]	Data Mining, Pattern Recognition, Feature Extraction.	To develop a method by using data mining techniques for the health-oriented database.	Well-defined explanation of data mining techniques from a health point of view with the help of three case studies.	No implementation is done.	Big data, High-speed computers, tablets.
Andre Goy, *et al.*[188]	Blockchain, Cloud, Robotics.	Using the latest technology introduced in the fourth industrial revolution provides the best and cheapest health services to all.	The latest technological usage is explained from a health point of view.	Time-taking process.	Mobile phones, biosensors, wearables, IoT.
W. Shi et al. [189]	Edge Computing, IoT.	A collaborative health environment is created where patients, pharmacies, hospitals, and government all are connected to this edge network.	Several cases of edge computing are defined with the help of its challenges.	No practical implementation is performed.	Mobile Phone.

Raffaele Gravina et al.[190]	Cloud Technology, Body Sensor Network (BSN), Cyber-physical Framework.	Full-fledged Cyber-Physical system is created for supporting on-line and off-line data information of human activity.	A framework is designed to provide 24/7 support to recognize and monitor human activities.	Data security issues are not handled properly.	Mobile Devices.
G. Sanninoa et al. [191]	Deep learning approach.	A novel deep learning approach is built for ECG classification.	A deep neural network-based approach is proposed for the automatic classification and differentiation of normal and abnormal ECG beats.	Real-world data processing capability is not introduced.	Electrodes on the body surface.
S. Sengupta *et al.* [192]	Cloud, IoT.	An e-Health framework was developed using cloudlet for real-time health data.	The framework improves the working of the conventional cloud-based framework.	The node failure issue and communication cost are not considered.	IoT devices, cloudlet.
C. Yvanoff-Frenchin *et al.*[193]	Mobile Applications, Artificial Intelligence, Edge Computing.	A multi-language robot interface is developed to evaluate the mental health of seniors by interaction through a set of questions.	A robot interface was developed that supports multi-language for evaluating seniors' mental health.	The robot can only be able to do its work in the presence of a specialist.	Mobile phones, Jupiter Notebook.

2.4.3 Identification of Research Gaps

- Most of the studies have used individual machine learning algorithms for disease prediction and they have not worked on different parameters of healthcare data management.
- Similarly, the studies have not worked on any particular algorithm that helps in specific disease detection.
- Very limited studies worked on the proper management of disease data, and not used architecture for healthcare data.
- Most of the studies focused only on disease data prediction using ML classification models and hybrid models. Limited studies focused on healthcare data management.
- There are several performance measures used for validating the algorithms. Most of the studies used only two or three performance measures for the validation process.
- Healthcare 4.0 approaches have not been used in the existing/available studies.

Following are the few research gaps that have been considered as the problem statement with their proposed solutions for this research work:

- In the absence of specific disease detection algorithm we developed an algorithm called MySymptom for CKD detection.
- To improve healthcare disease management problem, we use the multiphase architecture for the proper management of healthcare data.
- In addition to ML technique and hybrid models we also use parallel data distributed computing for data storage and management.

2.5 Summary

From the literature review, it has been established that there exists various research gaps in the area of healthcare data management, classification, specific disease detection, overall healthcare architecture, and H4.0 approaches as per the existing studies. Hence, there is pressing need of developing an efficient healthcare architecture to enhance the existing healthcare services, performance and reliability.

CHAPTER 3

Exploring Big Data for Healthcare

This chapter explains the essential components of healthcare big data such as clinical data, operational data, pathological test data, wearable device data, etc. The importance of BDA technology along with the associated challenges are also explained. The chapter presents the use of different tools for BDA and describes the technical details for five types of environmental diseases.

3.1 Introduction

The quality of life is directly proportional to the health of any person, which further depends on the quality of the environment. The rapid development of urbanization improved our daily life but also leads to a series of urban diseases. These prevalent urban diseases can cause cancer and many other environmental diseases in most Indian cities. Inadequate sanitation facility is the root cause of various environmental diseases. In the year 2016, hygiene, sanitation, and water were responsible for 829000 annual deaths due to diarrhoea, and this risk factor makes an important environmental contributor to ill health. Unsafe water, sanitation, or hygiene is the biggest factor in deaths around the globe due to diarrhoeal disease. In addition to diarrhoea various other diseases such as Malaria, Viral Hepatitis, Japanese Encephalitis, and Acute Respiratory Infection will be prevented if adequate water quantity is maintained with quality, hygiene, and sanitation facilities. Globally improving sanitation, hygiene and water facilities in developing countries has the potential to avoid at least 9.1% disease burden [194].

In the past few years, the quality of the environment deteriorated very fast and so is the case of health during a lifetime. A person suffers from various kinds of disease and visits government and non-government health centres. During this process, there is a generation of an enormous amount of unstructured data about the affected person. With the advancement of digitization technology, there is a substantial requirement for the documentation of this unstructured and inconsistent data and further this data should be analysed in a structured manner.

3.2 Inconsistent Data

The healthcare sector is a big contributor to the world's digital structured and unstructured data. Every second data is created from various sources. Usually, medical field data are produced from different sources and in different formats. Till now doctor's prescriptions are present in hard copy, and prescriptions may have some fuzzy contents in them based on the patient's reply. Because sometimes patients are not sure about their replies, what they gave to the doctor at the time of diagnosis. This problem leads to uncertainty in the diagnostic result, which is overcome by some pathological tests conducted to find out the exact problem. Now the reports of these tests are again in some different format which contains some abbreviations and a normal range to tally the results. The accumulation of this huge amount of data from different sources such as clinical data [195], pathological lab data, medical images, X-ray data, and sensor data has overburdened the capacity for effective and essential aggregation and analysis aiming to provide better clinical quality, personalized patient care, and patient safety. To figure out a decisive correlation or complex relations between the symptoms of a patient, diseases, and their respective treatment [196] one should do an intelligent and efficient data analysis representation. This efficient data analysis representation should also be able to provide better economical methods for cost reduction in personalized care. According to the World Health Organization (WHO), global health spending totalled $7.5 trillion in 2013, which is increasing by an average of 6% every year since 1995. This trend keeps on accelerating in the future and it reached $9.3 trillion in 2018 [197]. Although increased healthcare expenditure can be matched with the improvement in healthcare or the new and advanced services provided by it. These healthcare services can be differentiated between the following groups of our society.

3.2.1 Senior Citizens

Environmental health is the main component of the overall health of a Nation's citizens. Unfortunately, the division of environmental benefits and risks is not equal for all fragments of our society. Older people may have different vulnerabilities to environmental containments [198] due to various reasons like ageing parameters change in the degree of sustainability which is associated with physiological changes. Various other factors affect the vulnerability to environmental hazards factors such as socioeconomic and nutritional status [199]. Senior citizens require more medical facilities and care in comparison to the younger generation. This requires a big need for better and any time available healthcare medical facilities for them.

3.2.2 Citizens' Health

Nowadays the focus of the government is to provide better and safe health to its citizens. New healthcare-related policies are coming into existence and a large number of funds have been assigned to these policies. The Indian government has recently announced the world's largest healthcare program under the Ayushman Bharat program [200]. In this Program Policymakers and business practitioner around the world extend their extraordinary efforts in the field of e-health. Many local people awareness programs are also running which provide information to the citizens about infectious diseases like COVID-19, AIDS, swine flu, Malaria, etc. These awareness programs run under the supervision of local health authorities. Digitization makes the work easier to spread awareness among the citizens. The primary requirement of a government is to provide a healthy environment. But it is a challenging task for the government due to the contamination of the environment from the presence of different kinds of pollution. Industrialization is one of the major causes of it, which pollutes food, air, water, and soil from different pollutants. Even the radiations generated from the mobile towers are also harmful and due to lack of awareness in some places, they are placed near residential areas.

3.2.3 Increase in Chronic Disease

According to WHO air pollution contributes to a range of chronic diseases including asthma and other chronic respiratory diseases [201]. The main causes of chronic diseases are divided into various classes such as unhealthy diet, physical inactivity, intermediate risk factors (abnormal blood pressure), tobacco and overuse of pesticides, etc.

3.2.4 Government Policy and Mandates

To overcome the above-mentioned problems, in 2017 Indian government set various principles and objectives for national health policy (NHP). These are patient-centred & quality of care, professionalism, affordability, equity, integrity and ethics, dynamism and adaptiveness [202]. The objective is to provide free and basic primary health services to all pregnant ladies, children, and adolescents health to protect them from all communicable and non-communicable diseases present in the population. Government policy targets seven priority areas for improving the environment to provide better health.

- Reducing indoor and outdoor air pollution
- The Swachh Bharat Abhiyan (SBA)
- Provide balanced, healthy diets and regular exercises.
- Nirbhaya Nari (NN) –action against gender violence
- Yatri Suraksha (YS) – to prevent deaths due to rail and road traffic accidents
- Reduced stress for improving safety in the workplace
- Addressing tobacco, alcohol, and substance abuse

Swasth Nagrik Abhiyan (SNA) is a social movement to provide good health conditions to its citizens. Government policies are used for developing the institutional mechanisms and strategies for above mentioned seven priority areas to build a healthy living environment.

3.3 Big Data Analysis and Associated Challenges

Big Data Analytics is a new revolution in information technology. The healthcare sector is one of the most unexplored fields in terms of data analysis. As the data is accumulated from various sources such as wearable sensor technology, smartphone, healthcare apps, healthcare devices, etc. This data can be used for clinical predictions by using emerging machine learning algorithms and BDA techniques. BDA provides systematic information based on the vast amount of healthcare data to develop a better healthcare system. With the help of the analyzed patterns, valuable information can be extracted and used by policymakers to build a protective environment for a better healthcare system. For this, big data analytical tools are used but they are still in an incubator. BDA applications have the capability of improving healthcare results and reducing wastage of healthcare resources for the improvement of the overall value of healthcare. As an example, Google's model was formulated to predict the spread of influenza quite accurately in comparison to the US centre for disease control model, that is based on the submitted cases of health clinics and hospitals. The challenges of existing BDA technology are identified as follows:

3.3.1. Volume

It refers to the vast amount of data generation these days from various sources. According to reports, only U.S healthcare data reached 150 exabytes in 2011 and soon it will be going to reach a zettabyte in the near future if the growth rate remains the same [203, 204].

3.3.2. Velocity

It refers to the speedy generation of data from various sources which is making it big and complex to be handled by typical data handling software. In the medical field, this high-velocity data is generated from various sources such as wearable devices, smartphone applications data, sensors data, etc. To handle this tsunami of data powerful data engines are required. Probably, we can say Hadoop MapReduce helps for processing, storing, and retrieving this large amount of data.

3.3.3. Veracity

Since big data is a collection of a vast amount of data, so it is very tough to extract the genuineness of the data or any important information. This starts the journey of BDA because by applying it we can get valuable information based on related patterns. This can be done by using different tools.

3.3.4. Variety

Healthcare big data is a mixture of varied kinds of data which includes text and images from X-rays, MRI, sensors, radiology data, voice, video recordings, etc. The more variety of data can cause the chances of more errors. That’s why a relational database can no longer work on this mixed form of data.

3.3.5. Validity

The validity of data and veracity of data are two different concepts but very similar to each other. Generally, validity is used to check the accuracy and correctness of data on the behalf of its future usage. With the help of an example given by Huang et al. [205], the author explained that a physician cannot take data from a clinical trial related to a patient’s disease symptoms without proper validation. In other words, we can say that a set of data may be valid for one application but may be invalid for another usage.

3.3.6. Volatility

Big data volatility means the validity of data or how long the data is valid and how long it can be stored. In this, real-world time plays an important role as the data we need for a particular time may become futile for future analysis. In the medical field healthcare data changes according to time. This thing raised a question about the relevance of data that how long the data should be stored and when this data should be archived or deleted. As the size of data is increasing on daily basis it is an important decision to be taken by the experts.

3.3.7. Value

This is a special “V” among all Vs for a special reason. The desired outcome of the big data processing is retrieved by this “V”. Our primary goal is to extract the maximum value from any big data set.

3.4 Big Data in Healthcare

Big data usually is the vast amount of data whose size presents in terabytes, petabytes and zettabytes. Big data is growing so fast and it is having a complex nature due to its size. Although the concept of big data is not new, but the way it is defined is constantly changing [206]. Generally, big data is a mixture of various kinds of data elements whose size, type, complexity and velocity may vary. This type of composition helps in developing new business intelligence and helps in building new schemes for successful storage, data management, analysis and visualization of data elements [207]. Big data has the potential to spread its impact on various fragments of our society such as medicine manufacturing, economic policies, smart cities and healthcare.

In the health sector, the data is generally produced in a tremendous amount from different sources like city scans, MRI, pep tests, radiology tests, blood pressure monitory reports, wearable devices, fetal tests, etc. [208] as shown in Fig. 3.1. These scattered medical records are not sufficient to help the individuals at the time of major health problems. Hence, there is an urgent requirement to share this huge amount of healthcare data held by different stakeholders to improve the quality of this unstructured data. This huge amount of data has the potential to improve clinical decision support, population health management, and disease surveillance [209]. This data needs to be managed and analysed properly. The BDA is a step toward data management. For the success of the BDA initiative, the healthcare sector must do some fundamental changes regarding the maintenance of information in a digitized way by protecting patients’ privacy. This large and centralized healthcare data insist that healthcare organizations invest in BDA to gain valuable insight and knowledge from it and facilitates timely decision-making, minimizing patient risk, and reduces clinical cost [210].

To achieve these targets H4.0 approach helps the health sector as it is derived from the well-known I4.0 approach. H4.0 incorporates the principles of I4.0 for the digitization of laboratories and to implement automation in numerous processes used in the general health sector and hospitals [211]. H4.0 is used to improve the efficiency of physicians by enhancing their speed for exploring patients' data and, enabling them to optimize the resources so that patients' health can be improved. This all can be done by extending the boundaries of innovation with the collaboration of the internet of services and the internet of things (IOT) [212].

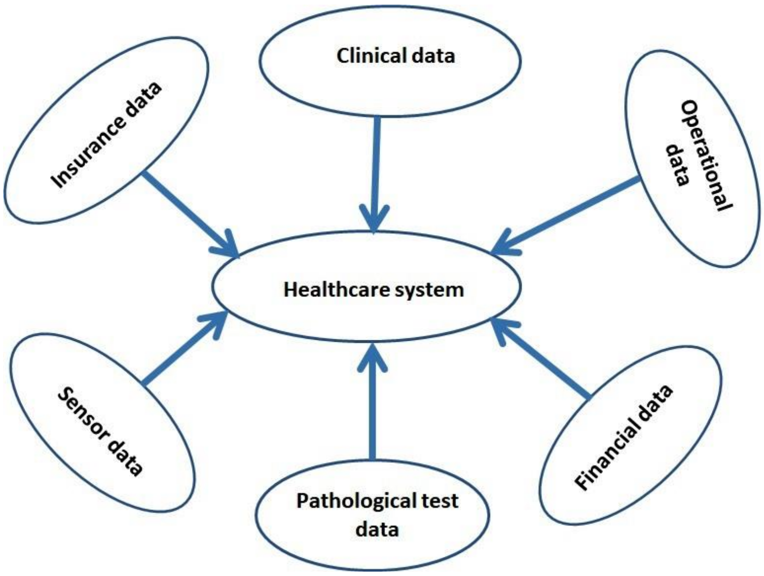


Figure 3.1: Representation of Different Sources of Data in the Healthcare System.

3.5 Big Data Analysis Using Different Tools

Big data sets of healthcare refer to the large electronic health data set. The management of these large electronic data sets is very tough to use in a traditional way or from traditional tools. Traditionally, business intelligence tools and applications of stand-alone systems along with limited clinical data processing ability are used for healthcare data management. Now new tools with high processing capabilities are on the market, a few of them are listed in Table 3.1. The storage of huge volume and unstructured data is a big task which is solved by Apache HBase, NoSQL, and Hadoop.

Hadoop is an open-source software supported by the Apache software foundation. Hadoop works on commodity hardware and supports distributed storage and processing of an immense amount of data. Hadoop framework consists of MapReduce, HDFS, and Yet Another Resource Negotiator (YARN). HDFS is written in Java and supports master-slave architecture.

Here, the master node is performing the tasks of the NameNode and client nodes are performing the task of the DataNode. There is only one NameNode per cluster and several DataNodes. HDFS provides a robust environment by maintaining replicas, and it also maintains a secondary NameNode for emergency conditions, when the primary NameNode stop working. The NameNode and DataNode communicate through heartbeat messages. Hadoop plays a very important role in the processing of this large amount of data and extracting relevant information.

Table 3.1: List of Big Data Analytics Tools with Their Uses

Name of the tool	Used for
HDFS	It provides distributed storage of data on a cluster. HDFS distributes the data into small blocks and stores it across the cluster.
MapReduce	MapReduce is used for the parallel processing of applications.
Pig	Pig is a tool available in Hadoop to substitute the ETL process. It works in a language known as PigLatine.
Hive	Hive works on HQL language which is similar to SQL. It works on the data sets stored in Hadoop but data should be present in some sorted order or in a structured form.
Strom	Strom is used for real-time data analysis. It takes data from many data sources such as Twitter, ActiveMQ, Kafka etc. Strom converts the data into a manageable form and stores it on any database such as SQL or NO SQL [213].
HBase	It works on a column-based concept to store the data in a distributed database system. HBase works on top of HDFS. A real-time query can be generated in the Hadoop environment by using HBase.
Solar	Solar is used to perform searching in text documents. It maintains indexes and performs searching based on keywords.

3.6 Healthcare Ecosystem

The recent advancement toward digitization in the healthcare industry has opened up new research opportunities. Information technology has already been used in the healthcare sector in many forms. Following are a few case studies available in support of the above statement. Global Positioning System (GPS) enabled trackers are used to monitor the usage of the inhaler by asthmatic patients. This information is centrally stored in a database to identify individual group or person-based trends. It helps the physicians to develop personalized treatment plans for the patient so that preventions can be taken in advance. Some mobile applications are also assisting in analysing behavioural health therapies. With the help of mobile sensors present in smartphones, the healthcare application records the different activities of a particular person. Based on the obtained insights from the healthcare applications, the person's behaviour can be analysed. It figures out the physical activity of a person such as working hours, irregular sleep patterns, anxiety level, etc. and predicts the health condition of that person.

To reduce the financial burden of the health system for end-stage renal disease (ESRD), replacement therapies accounted for more than 2.5 billion euros annually [214]. A kidney transplant is done for improving the quality of life of a person, but the survival rate after a transplant may depend on the recovery of a person. The primary goal is to reduce the complications and increase the lifespan of a person by giving proper care to deal with life-threatening complications such as severe infection, and acute rejection malignancy. The important factor is to figure out some chronic infections and side effects such as cardiovascular problems, costly medications, and hospitalizations. [215, 216, 217, 218, 219].

To perform the above-mentioned tasks the healthcare data should be maintained in a proper way so that meaningful and relevant information can be extracted. This information is further used for research purposes and can be utilized to create new trends. All of this comes under the BDA process that can potentially be used to create a good health ecosystem by accessing large and diverse data sources to provide a timely estimation of quality.

BDA was used at Columbia University medical centre to analyse the complex correlation between the physiological data of a patient with the record of its brain injuries. The goal of this study is to provide timely information to medical professionals for treating critical complications [220]. In a similar way in California, a partnership project was conducted by blue shield and Nant health for advancing care delivery and improving outcomes. The developed outcome has been proposed based on an integrated technology system which may allow hospitals, doctors, and health planners to deliver evidence-based healthcare data. It would be more coordinated and personalized for improving the prevention and care coordination to promote the right care pathway.

Moreover, early disease prediction is beneficial to provide good and stable health. For this, information technology, H4.0 and BDA boost the process of data maintenance and data analysis in a proper way. This makes a huge amount of healthcare data easier to store and process and extract some related patterns and valuable information. This may be used by every individual of the dependency chain such as policymakers→ hospitals→ Pharmaceutical companies→doctors→ patients associated with each other. BDA technology and H4.0 is in an incubator state but business values extracted from these methods definitely provide help to any particular entity of this dependency chain to accelerate their maturing process. BDA can potentially be used to create a good health ecosystem by accessing large and diverse data sources to provide a timely estimation of quality.

3.7 Experimental Method

The experimental method involves the empirical analysis of five diseases that are caused due to adverse environmental conditions and the practice of poor hygiene. Globally diarrhoea is the second most important cause of death in children under five years. The main causes of these diseases are a variety of bacterial, viral, and parasitic organisms which leads to infection in the intestinal tract. The infection spreads through polluted food or drinking water or as a result of poor hygiene. These diseases can be prevented by using safe drinking water or by using improved sanitation practices to reduce disease risk. The flowchart of Fig. 3.2 explains the overall working of the experimental model with the help of different steps.

Step 1: In this step, we are retrieving the five environmental disease datasets available on the Indian Government website [221]. The data at the website “data.gov.in” contains data related to different diseases in a vast amount, it also contains the data of sanitation and water facility in India along with the health resources.

Step 2: In step two, pre-processing of the dataset is performed by normalization. It is very tedious work to perform this step as the amount of data is huge and it consists of different disease data along with various causes of disease. However, the implementation of big data in healthcare is in an incubator stage but can be used to predict some significant information.

Step 3: In this step, dataset is passed to all classifiers individually by applying machine learning techniques. The result generated by these classifiers will help to predict the disease.

Step 4: Here, Recall and Precision performance measures are calculated to measure the performance of all the models.

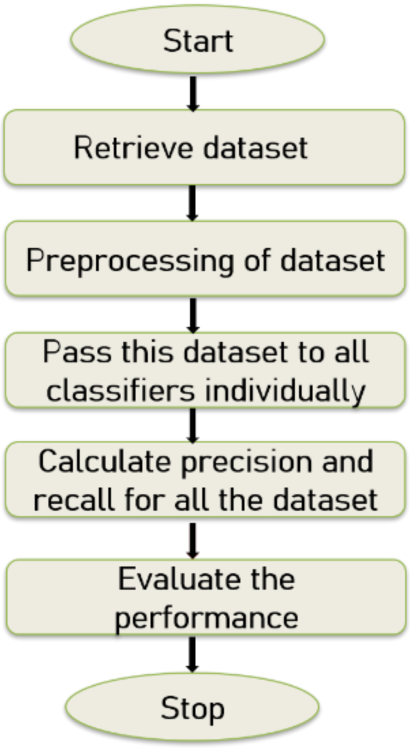

Start
Retrieve dataset
Preprocessing of dataset
Pass this dataset to all classifiers individually
Calculate precision and recall for all the dataset
Evaluate the performance
Stop

Figure 3.2: Representation of the Working of an Experimental Model

3.7.1 Results and Discussion

The Recall values of different models for the various diseases are tabulated in Table 3.2 for different data sets. The validation of results is done by the 10-fold cross-validation method. The results generated by precision and recall are considered best when their value is close to one. Here, it is observed that for recall NB classifier produces the best results with Viral Hepatitis, Malaria, and Japanese Encephalitis. This shows that Malaria, Viral Hepatitis, and Japanese Encephalitis have more independent attributes. For Diarrhoea, RF classifier performs well and the Logistic Regression classifier works for Acute Respiratory. Figure 3.3 represents a bar graph for various models and is plotted using the data from TABLE 3.2. Hence, the observed result indicates that the best model is created for Viral Hepatitis with NB and it gives a maximum 97% value among all the diseases for Recall.

Table 3.2: Calculated Recall for Various Models Using Classification Techniques

Diseases data set	**Support Vector Machine**	**Naïve Bayes**	**Random Forest**	**Decision Tree**	**Logistic Regression**
Malaria	0.36	0.64	0.41	0.52	0.26
Diarrhoea	0.75	0.50	0.76	0.70	0.66
Viral Hepatitis	0.31	0.97	0.30	0.42	0.14
Japanese Encephalitis	0.21	0.45	0.35	0.42	0.17
Acute Respiratory Infection	0.77	0.54	0.85	0.81	0.87

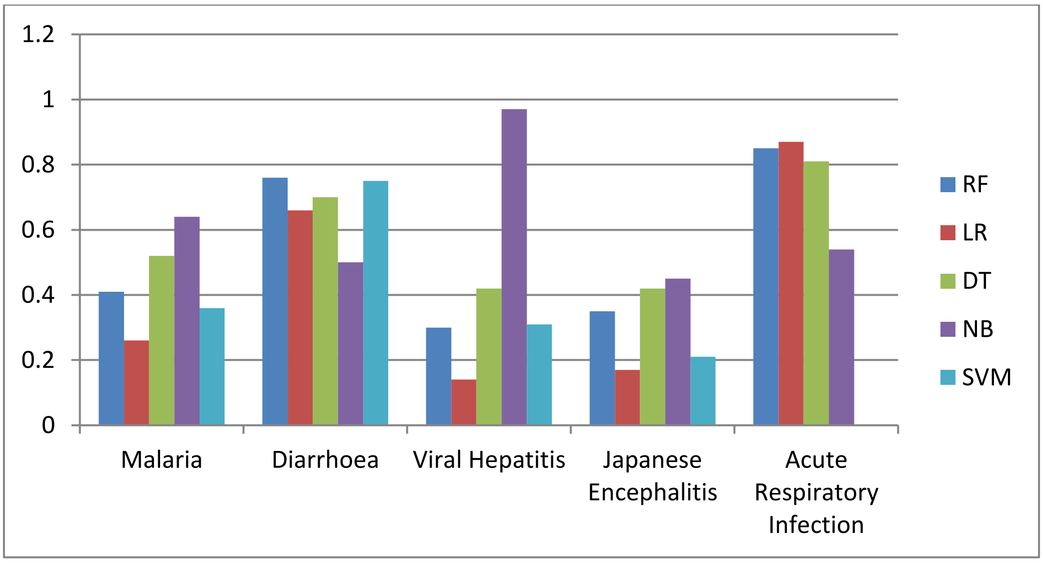


Figure 3.3: Graph of Calculated Recall as Performance Measures

The Precision values of different models for the different disease is calculated and represented in Table 3.3. The obtained value of precision for various ML models was calculated on different environmental disease datasets. The best precision is measured by the RF classifier in the case of Malaria, Viral Hepatitis, and Japanese Encephalitis. For Acute Respiratory Infection RF classifier and NB classifier both produce the same results. As in the Diarrhoea dataset, NB performs to give the best results. This means that the Diarrhoea dataset has more independent attributes than any other dataset. Precision and Recall is a very important parameters required for measuring the performance of a classifier. Performance measures like Precision and Recall are significant factors in the medical field as the concept of promise to achieve a better healthcare system with a more proactive to predictive approach is achieved by applying these measures. Figure 3.4 represents a bar graph for various ML models and is plotted using the data from Table 3.3. Thus, the observed results predict that the best model is created for Acute Respiratory Infection with RF and Naïve Bayes classifiers and it gives 97% value among all the diseases for Precision.

Table 3.3: Calculated Precision for Various ML Models Using Classification Techniques

Diseases data set	**Support Vector Machine**	**Naïve Bayes**	**Random Forest**	**Decision Tree**	**Logistic Regression**
Malaria	0.49	0.46	0.56	0.44	0.45
Diarrhoea	0.72	0.84	0.77	0.71	0.74
Viral Hepatitis	0.25	0.30	0.56	0.40	0.41
Japanese Encephalitis	0.41	0.41	0.55	0.42	0.44
Acute Respiratory Infection	0.83	0.97	0.97	0.94	0.96

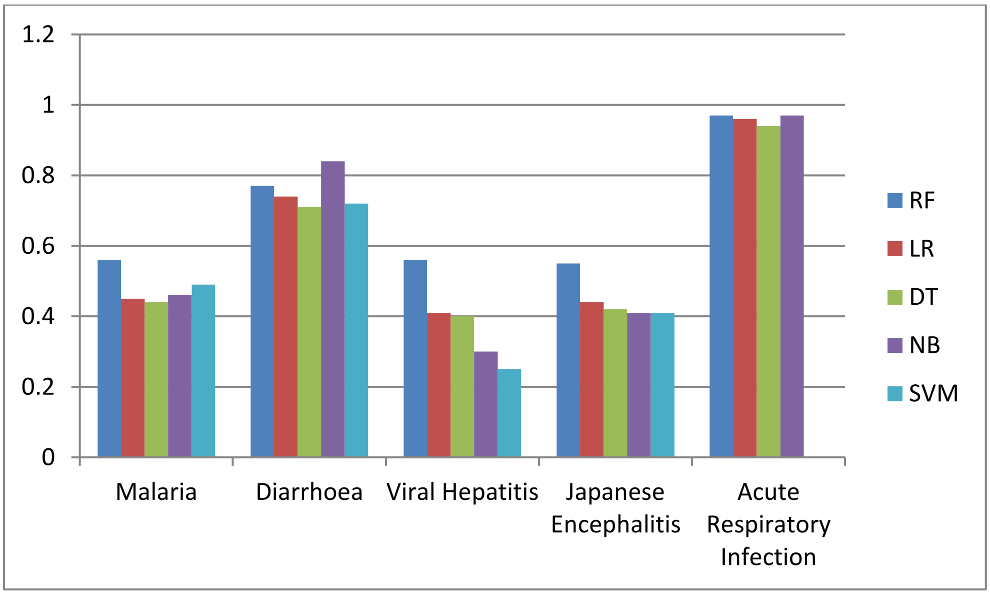
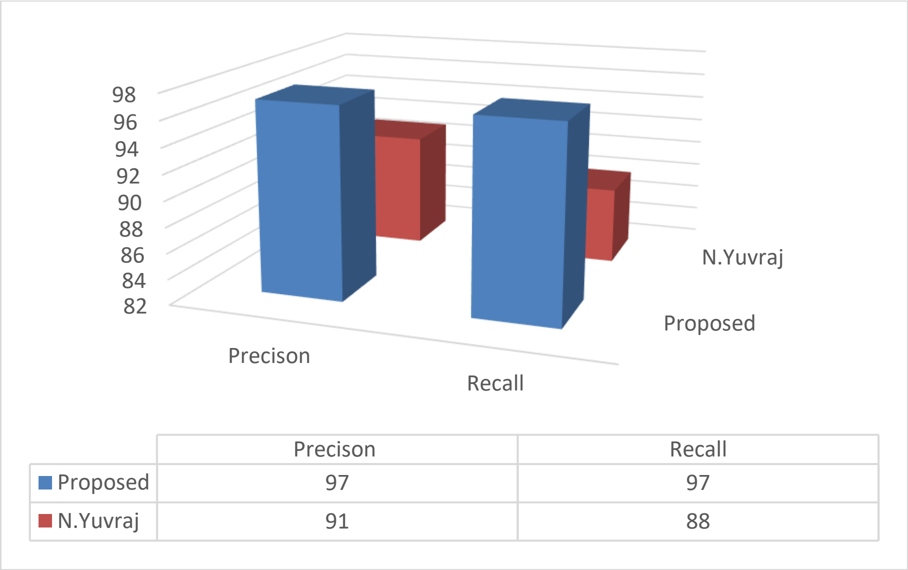


Figure 3.4: Representation of Precision for Various Models

3.7.2 Comparison with Other Techniques

The investigation carried out by N. Yuvaraj et al. [108] uses a diabetes data set in a Hadoop environment and performs predictive analysis by using three ML algorithms and comparative results with the present study are shown in Fig.3.5. In their analysis, NB produces the best precision value of 91% and RF produces best recall value of 88%. In the proposed method we produce our result by using five ML algorithms on five different environmental disease data sets. The best precision value produced by NB and RF for Acute Respiratory Infection is 97% and the best recall value produced by NB for Viral Hepatitis is also 97%. In both cases, the result of the proposed method is better than the method used in the comparative study [108].



	Precison	Recall
🟦 Proposed	97	97
🟥 N.Yuvraj	91	88

Figure 3.5: Comparison of the Results of The Proposed Study with N.Yuvraj et al. [108]

3.8 Summary

The BDA technology and associated challenges are explained in the chapter along with the tools to process healthcare big data. This chapter predicts the particular environmental diseases using a 10-fold cross-validation technique for an unbiased result. Among the data set of different diseases such as Malaria, Diarrhoea, Viral Hepatitis, Japanese Encephalitis, and cute Reparatory Infection the best Recall and Precision values are observed to be 97% for Viral Hepatitis and Acute Respiratory Infection with RF and NB classifiers respectively. The proposed model is also compared with the state-of-the-art algorithms and it gave the best precision and recall values. Moreover, it has been further improved by performing proper healthcare data management and analysis of disease based on its symptoms in the subsequent chapters with the help of healthcare architecture.

CHAPTER 4

Healthcare Multi-Phase Architecture for Disease Prediction

It has been observed from the earlier discussions that there is an utmost requirement for a healthcare architecture that manages healthcare data and at the same time it should also be able to identify the disease based on its symptoms so that it can be easily analysed. This chapter proposes an HCMP architecture that handles all the issues for predicting CKD.

The HCMP architecture works on six different layers where data-storage and data-management layers were performed on heterogeneous Hadoop cluster and the profiling methods were used to consider three situations for calculating the capacity ratio of each DataNode in the cluster. MySymptom algorithm has been used to filter the kidney dataset of patients according to their symptoms at the data processing layer. In addition to this horizontal scaling is performed in the Hadoop cluster, and the performance of every DataNode is investigated based on a capacity ratio. The data-analysis layer has performed classification tasks using different machine learning algorithms.

4.1 Introduction

In recent years, due to the advancement in information technology data were generated exponentially. The digitization of the healthcare sector makes it a complex enterprise. The arrangement of healthcare data should be done in a proper way to extract the important information. Big data analytics is used to uncover new insights and hidden patterns out of this digitized data. This will help the doctors to take the right decision at right time. The refined information is useful in providing a better clinical decision support system (CDSS). Therefore, the complete information about patients will be very helpful for good statistical analysis, data mining, and Machine Learning (ML) [222].

The analysis of healthcare big data and human interpretation is a challenging task as data is present in an unstructured format [223]. ML algorithms and artificial intelligence are massively applied to extract vital information from data. ML algorithm uses a large dataset to learn and predict the future outcome. It also helps in the management of chronic disease by doing timely calculations.

Nowadays, CKD is a critical health problem. The subtle nature of CKD does not allow it to identify at an early stage [224]. CKD is not showing any dominant symptom at an earlier stage and it is very difficult to identify CKD without any urine or blood test. The identification of disease at an earlier stage helps to provide better treatment at right time and this reduces the chances of dialysis and transplantation. But it is an impossible task to check the possibility of CKD in all persons as it will go under an extensive testing process. At present data preservation is also a tough task in the healthcare industry and it is very important to perform a classification task [225].

For data management, Hadoop clusters are used. Generally, the Hadoop cluster is being used to store, manage and process a huge amount of unstructured data. The Hadoop data cluster distributes the workload among multiple DataNodes that work in parallel. Once the data is organized in a proper manner it can be utilized for further analysis and the value of the data can be extracted. Value is the soul of any data and it helps to understand the three W’s of the data i.e. what, when, and why [226]. Here, what indicates the type of data, why tell about the requirement of data and finally when indicates the use of the data.

In this investigation, we are trying to adhere the well-established eminent requirements and applications of health informatics with the help of five Es of journalistic reporting [227]. Further, these five Es were correlated with the various layers of proposed HCMP architecture and explained as;

- Experimental: Experimenting the things before reaching any conclusion is a prerequisite for every study. With the help of the proposed architecture doctors can diagnose the disease of the patients based on their symptoms and similar history patterns. MySymptom algorithm is used to assign the score to the patients and this score depends upon the similarity of the patient’s attributes.
- Experiential: Experience comes with time and in a scenario where changes occur so fast, the lifestyle of people is changing and the environmental conditions are also

behaving abnormally. Due to all of these changes a new type of disease came into existence. The requirement for a new type of study and new methods in the medical field is becoming a crucial urge. There is an urgent need to introduce a new technology that helps health practitioners and government officials to provide the right information on time for taking crucial decisions. This thing can be attained with the help of different machine learning algorithms.

- Explanatory: Explaining things properly to patients before reaching any decision is the basic duty of every medical practitioner. Evidence requirement is the primary need behind every decision to explain the exact condition of the patient. Hence, health practitioners required pathological test reports and based on the report they can reach any conclusion. The collection of different lab reports of different patients is done to extract out the similarity between them so that personalized health services can be provided to them.
- Emotional: Emotional journalism stands on the other side of the spectrum. With the advancement of technology, people can communicate with each other very easily and share reviews of healthcare facilities. These reviews help the government officials for knowing the ground reality of a particular area's healthcare facilities.
- Economical: Budget constraining is the biggest problem faced by every industry, whether it is the health industry or journalism. To attain experimental, experiential, explanatory, or emotional things within a fixed period budgeting is a big task. With the help of big data analytics and ML, the budgeting can be controlled as it provides fast results with minimum resource utilization.

This shows a major technological challenge to develop a unified and smart framework for healthcare. In this chapter, a disease chain management plan is maintained, and based on the observation it shows that disease does not happen in isolation instead of this it’s a consequence of a related occurrence. It happens with the interaction of various events such as food habits, lifestyle, environmental effects, and genetic factors. Patients who share the common risk, lifestyle habits, and environmental conditions may have similar risk patterns or outcomes. Big data provides an advanced learning and training environment for calculating disease prediction factors among patients. The proposed HCMP architecture consists of six different layers and helps to fulfil the above-mentioned requirements. All layers of architecture are dedicated to handling various issues.

4.2 Proposed Work

This section presents the entire proposed work in detail.

4.2.1 HCMP Architecture

Health care multi-phase architecture consists of six layers as shown in Fig. 4.1. The HCMP architecture deals with various challenges such as capturing large-scale healthcare data that is present in a mixed format. Secondly combining different technologies to produce a unanimous result for the betterment of society. HCMP architecture journey starts with data collection from various sources, then storage of data and management of data was done on Hadoop cluster that uses parallel processing. Analysis of data is done by RapidMiner by using the Radoop extension. Different ML algorithms are used for CKD prediction. In the end, report generation is performed for various stakeholders. The different layers of HCMP architecture are:

- Data collection layer (DCL)
- Data storage layer (DSL)
- Data management layer (DML)
- Data processing layer (DPL)
- Data analysis layer (DAL)
- Report generation layer (RGL)

4.2.1.1 Data Collection Layer (DCL)

The work of the DCL is to capture the data from various sources such as hospital data, patient’s demographics, sensor data, environmental data, doctor’s prescription, pharmacy data, geographical data of that place, medical insurance company data, etc. All the collection part of data is done in this layer at a central place. The data is present in mixed format and it will be handled by a new layer very efficiently for data storage and data management.

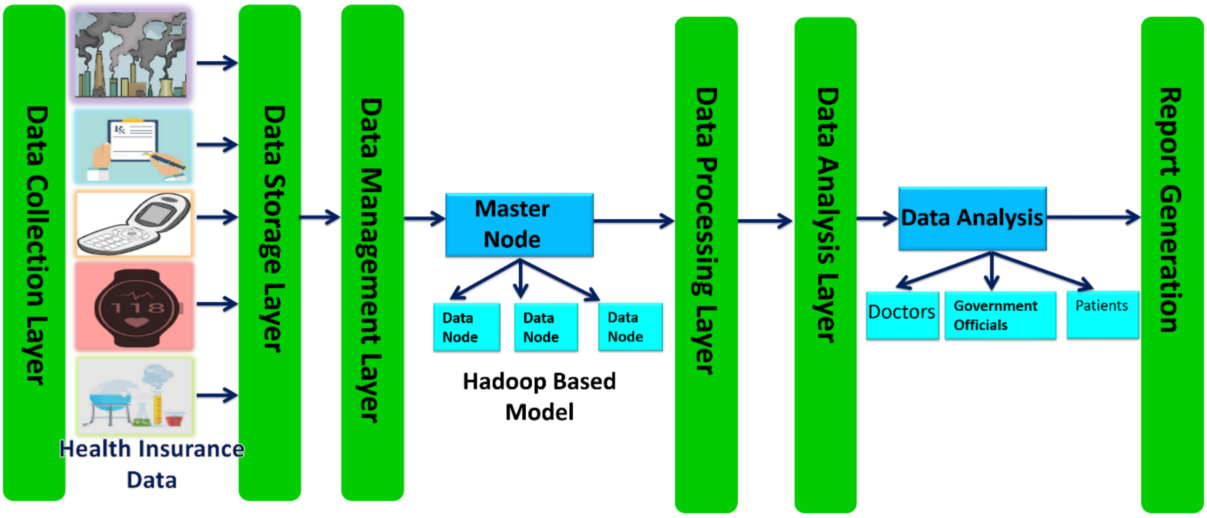


Figure 4.1: The Working Model of HCMP Architecture in Its Six Layers

4.2.1.2 Data Storage Layer (DSL)

To store large size of healthcare data Hadoop Distributed File System (HDFS) is used. The storage component of Hadoop is HDFS [228]. HDFS gives the best results to deal with large files for example reading and writing the file whose size lies in hundreds of megabytes (MBs), gigabytes (GBs), or terabytes (TBs) in size. The cluster of HDFS is a combination of a NameNode and several DataNodes. The NameNode acts as a master node and takes important decisions to manage the entire file system namespace. NameNode maintains a proper record of all file systems and the metadata of all the files and directories are maintained on a tree in proper order [229]. DataNodes work as slave nodes in the HDFS cluster. The DataNodes are used to store and retrieve data blocks after receiving the instruction from NameNode. Then DataNode informs back to NameNode regularly with the name of blocks present in its local memory. NameNode is the only access point to operate the cluster, without NameNode the HDFS file system cannot be accessed. In the Hadoop cluster, MapReduce is used for parallel processing to store the structured data on HDFS. MapReduce is the programming framework at the heart of Apache Hadoop. It is utilized to provide scalability to Hadoop clusters with thousands of DataNodes. MapReduce is a programming model used for processing massive volume of data across several clusters. In the MapReduce job processing, there are two crucial phases: Map and Reduce. In the file system, each phase has input and output key values. MapReduce is a data-intensive parallel computation framework that enables high performance through parallel data processing. The MapReduce task runs on unstructured data to organize it according to the user's requirements. The structured data will then be properly saved on HDFS for subsequent processing.

4.2.1.3 Data Management Layer (DML)

The proper management of healthcare big data is the leading problem faced by healthcare organizations. For managing the huge amount of healthcare big data a distributed data management technique is required [230]. Hadoop works on commodity architecture and with the help of MapReduce, HDFS provides a safe mechanism for storing and retrieving a large amount of data by replicating it on various stacks. The default architecture of Hadoop works on master-slave mechanism. Hadoop considers each DataNode equally in its default environment but in a heterogeneous cluster, every DataNode is different based on its configuration. A high-speed DataNode can perform its task of data processing speedily presents on its local disk in comparison to data presents on a slow node’s local disk. According to Hadoops default strategy, a fast DataNode after completing its local processing helps one of the remote slow DataNode to process its unprocessed data. This transferring of data from the slow DataNode to a fast DataNode affects the performance of the Hadoop cluster. It makes the master node very busy in maintaining the blocks and nodes information of a cluster. The master node has to manage all the statistical information about a node and the jobs running on it. NameNode calculates the total block number by dividing the data size (N) by the block size (M). In the proposed study the block size is set as 128MB. The NameNode also maintains proper information set regarding the empty task slot, unprocessed data, and a way to allocate this unprocessed data to a DataNode with an empty task slot. Hence, for handling this heterogeneous behavior of DataNodes an efficient data placement policy is required. This problem is handled by the proposed algorithm that distributes the data among different DataNodes based on their performance after calculating the Capacity Ratio (CR) for each DataNode. According to this a DataNode whose performance is better will get more data blocks to process and the slower DataNode gets fewer data blocks. This thing can control the migration of a vast amount of data on the network from slow DataNode to fast DataNode. As the network is a scarce resource and migration of a vast amount of data creates congestion in the network that may deteriorate the system and Hadoop performance.

The working of the Hadoop model is divided into three phases. Working of the first phase shown in Fig. 4.2 and Algorithm 4.1. Initially, NameNode does not have any information about the capacity of DataNode for various types of jobs.

In the beginning, when a job came for execution NameNode distributes the data among the DataNodes equally [231]. Therefore, after executing the particular job type each DataNode returns its task execution time to NameNode through heartbeat messages. NameNode performs the calculation task by calculating the CR of DataNodes and maintains all information in a CR table. The same procedure will be applied in the future if the same type of job arrives for execution and at that time NameNode distributes the data among the DataNodes according to their CR. Algorithm 4.1 explains the working of phase one and it calculates the total number of data blocks before distributing them to DataNode. After calculating the total block number each DataNode will get the data blocks according to their capability obtained from CR.

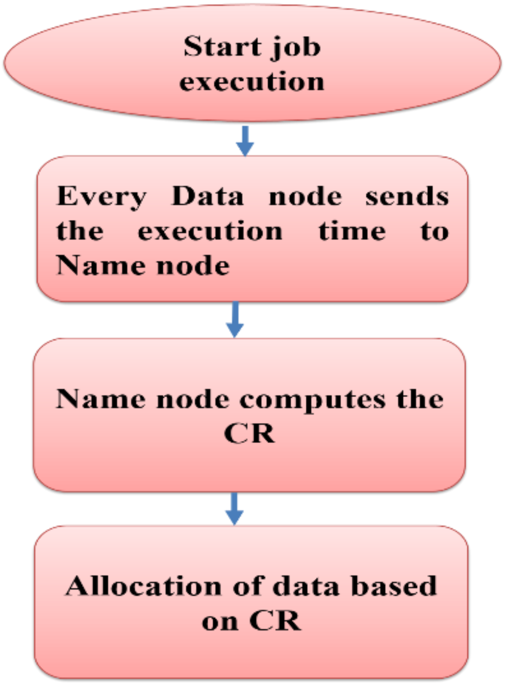

Start job execution
Every Data node sends the execution time to Name node
Name node computes the CR
Allocation of data based on CR

Figure 4.2: Phase One of Data Distribution

```
Algorithm 4.1: Initial_data_placement.
When data is first time written on HDFS:
JobType ← type of job performed by the DataNodes;
for the first time data is distributed in the ratio of 1:1:1;
Data size $(N)$ ← acquire from data information;
Block size $(M)$ ← 128 MB;
Total block number= $\left\lceil \frac{N}{M} \right\rceil$;
for all records in the ratio table do;
    Set the performance time according to job type in the ratio table for all DataNodes and jobs do;
    Node capacity ← obtain from the capacity ratio;
    Block number= Total block number $* \left[ \frac{Node\ capacity}{\sum each\ node\ capacity} \right]$;
    Assign the block number of data according to the performance to respective DataNodes.
```

The working of phase two is shown in Fig. 4.3 and Algorithm 4.2. This phase starts when the value stored in CR is changed or a new job came for execution. Whenever a new job came the NameNode checks the ratio table created in phase one. If the job type is present, then the NameNode distributes the data blocks among the DataNodes in the same proportion to process the same job type. At this point, no data blocks are required to transfer from fast speed DataNode to slow speed DataNode. Here, if the type of job is not present in the CR table then NameNode has to calculate the CR for each DataNode based on response time. After this, reallocation of data blocks in the DataNodes is performed based on the new CR. Then finally NameNode modifies the CR in the table for future use.

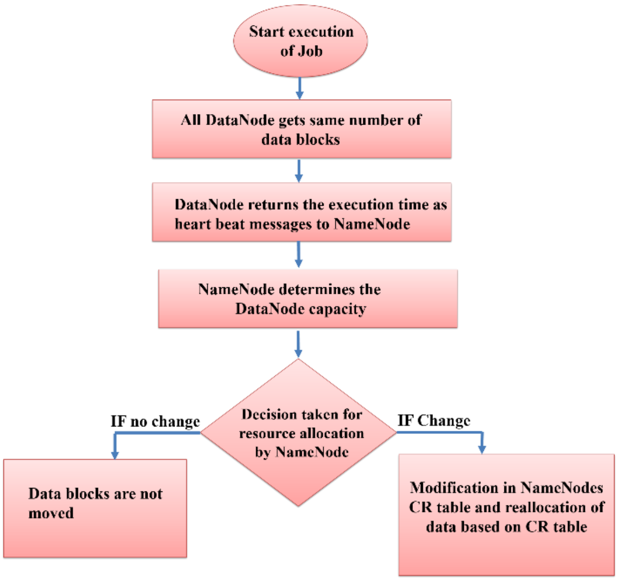
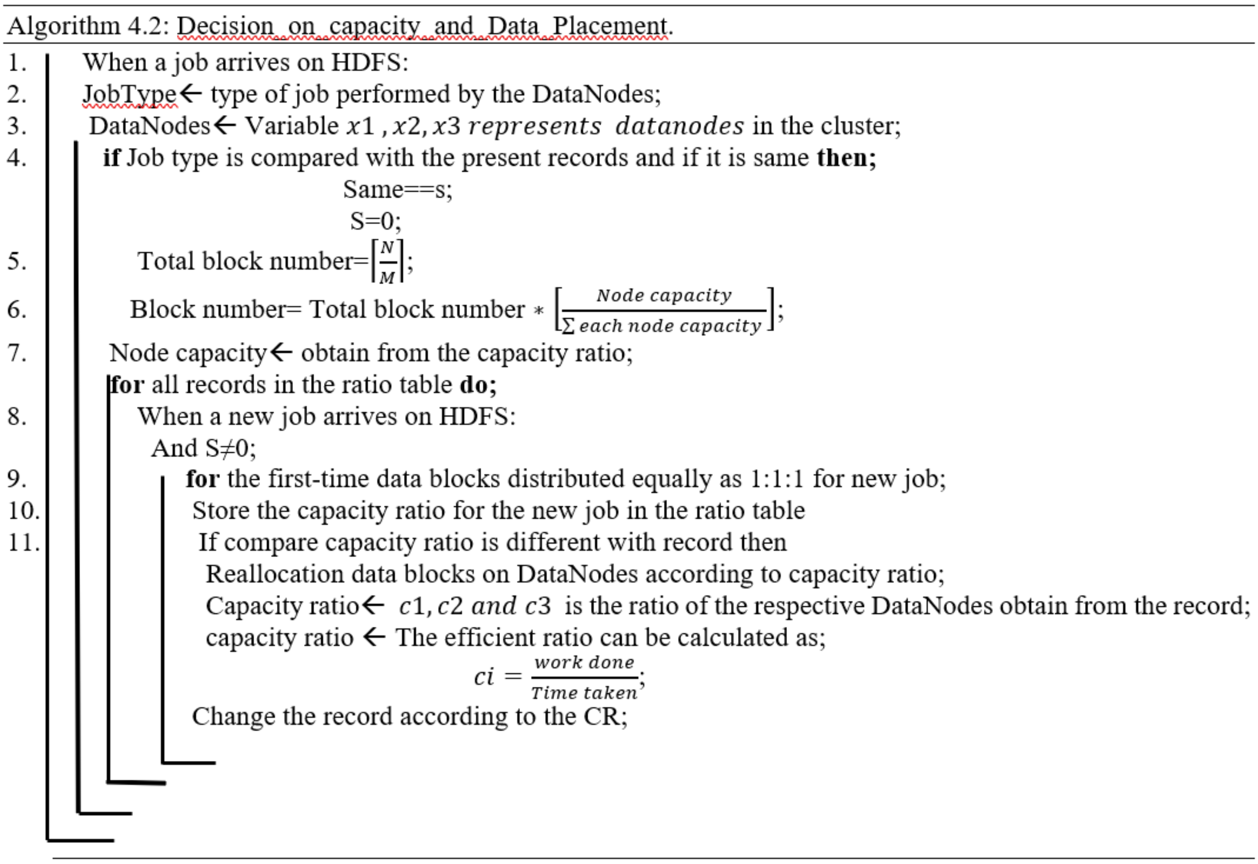

Start execution of Job
All DataNode gets same number of data blocks
DataNode returns the execution time as heart beat messages to NameNode
NameNode determines the DataNode capacity
IF no change
Decision taken for resource allocation by NameNode
IF Change
Data blocks are not moved
Modification in NameNodes CR table and reallocation of data based on CR table

Figure 4.3: Working of The Second Phase for Data Distribution



```
Algorithm 4.2: Decision on capacity and Data Placement.
When a job arrives on HDFS:
JobType← type of job performed by the DataNodes;
DataNodes← Variable $x1, x2, x3$ represents datanodes in the cluster;
    if Job type is compared with the present records and if it is same then;
                        Same==s;
                        S=0;
        Total block number=$\left[\frac{N}{M}\right]$;
        Block number= Total block number $* \left[\frac{Node\ capacity}{\sum each\ node\ capacity}\right]$;
      Node capacity← obtain from the capacity ratio;
      for all records in the ratio table do;
        When a new job arrives on HDFS:
          And S≠0;
            for the first-time data blocks distributed equally as 1:1:1 for new job;
            Store the capacity ratio for the new job in the ratio table
            If compare capacity ratio is different with record then
              Reallocation data blocks on DataNodes according to capacity ratio;
              Capacity ratio← $c1, c2\ and\ c3$ is the ratio of the respective DataNodes obtain from the record;
              capacity ratio ← The efficient ratio can be calculated as;
                              $ci = \frac{work\ done}{Time\ taken}$;
            Change the record according to the CR;
```

The third phase is used to improve the scalability of the Hadoop cluster by performing horizontal scaling. The working of the third phase starts at the time of a new DataNode's admission in the existing cluster. When a new DataNode is admitted in the cluster the NameNode is not having any log for the new DataNode. At this time the ratio table of the NameNode will become futile. To overcome this problem NameNode makes a log for new DataNode in its table by making a new entry. The first time after adding this new DataNode, NameNode distributes the minimum amount of data blocks to the newly admitted DataNode. In the experiment, for the first time 10% value is fixed for the new DataNode and the remaining DataNodes get 90% of data based on their CR. This technique is used to improve the efficiency of the overall system. After this step NameNode calculates the CR of the new DataNode based on the response given through heartbeat messages. Now NameNode modifies the records in the ratio table and makes the actual entry of CR for the new DataNode. Figure 4.4 shows the working of the third phase where NameNode makes a judgment depending on the cluster's updated resources.

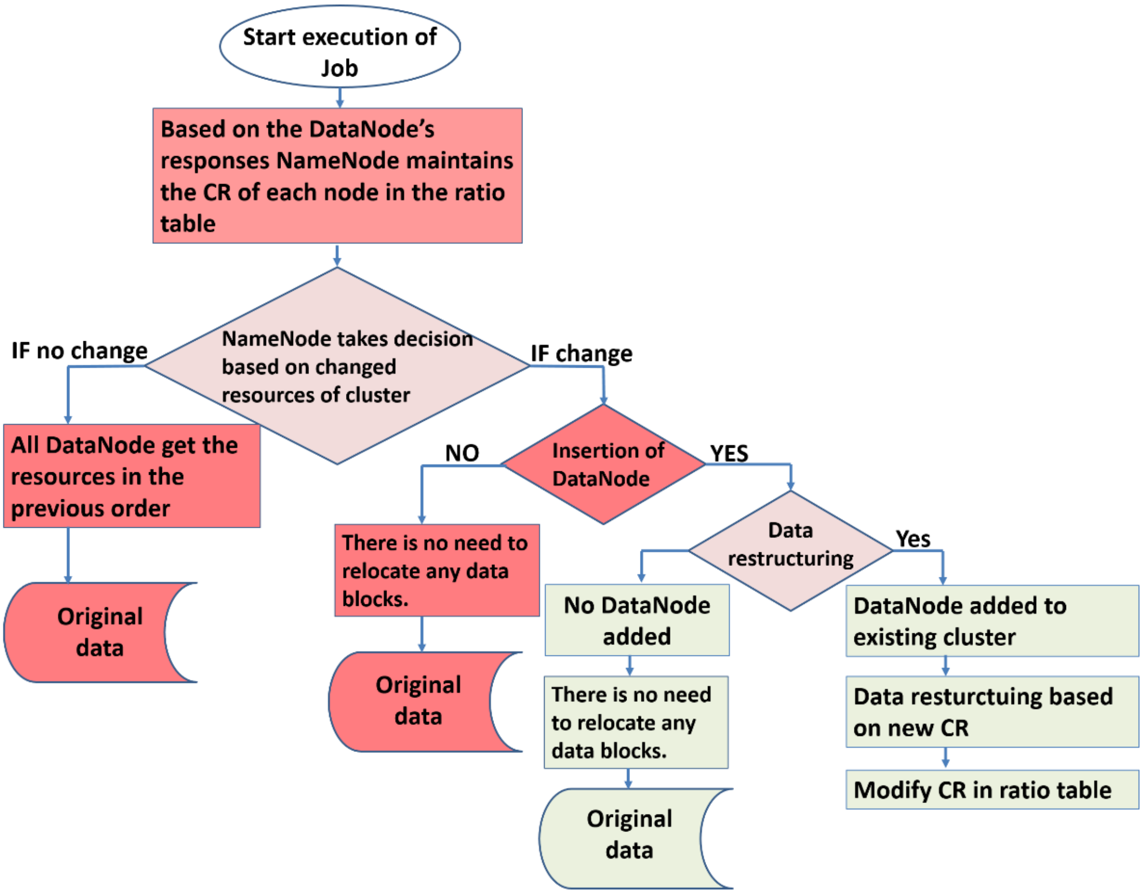

Start execution of Job
Based on the DataNode's responses NameNode maintains the CR of each node in the ratio table
IF no change
NameNode takes decision based on changed resources of cluster
IF change
All DataNode get the resources in the previous order
Original data
NO
Insertion of DataNode
YES
There is no need to relocate any data blocks.
Original data
Data restructuring
Yes
No DataNode added
DataNode added to existing cluster
There is no need to relocate any data blocks.
Data resturctuing based on new CR
Original data
Modify CR in ratio table

Figure 4.4: Working of the Third Phase, When a New DataNode is Admitted in the Existing Cluster

4.2.1.4 Data Processing Layer (DPL)

The data processing step is performed to remove the noise, unwanted data and conversion of categorical values into numerical values. In this model for data processing MySymptom algorithm is used. MySymptom algorithm ignores the missing values in the database and uses collaborative filtering. Collaborative filtering is an approach that is used to identify similar individuals through a set of common symptoms or attributes shared by a group of people [232, 233, 234]. As shown in Fig. 4.5, the MySymptom algorithm is divided into three major steps. The working of first step starts when an individual arrives at a doctor's clinic with a specific problem like CKD. In the next step, the medical history of a patient will be compared with the record of other patients' who may share similar symptoms. The basis of these similarity constraints are disease symptoms, family history, similar lab reports, the same environmental conditions, stress level, patients' demographics, Diabetes mellitus, etc. In the last step, scores have been assigned to the patients based on their disease symptoms similarity and the scoring ranges between 0 to 1. A score near 1 shows that the patient may suffer from CKD while a score near 0 indicates that he will not going to have CKD in the future as shown in Table 4.1.

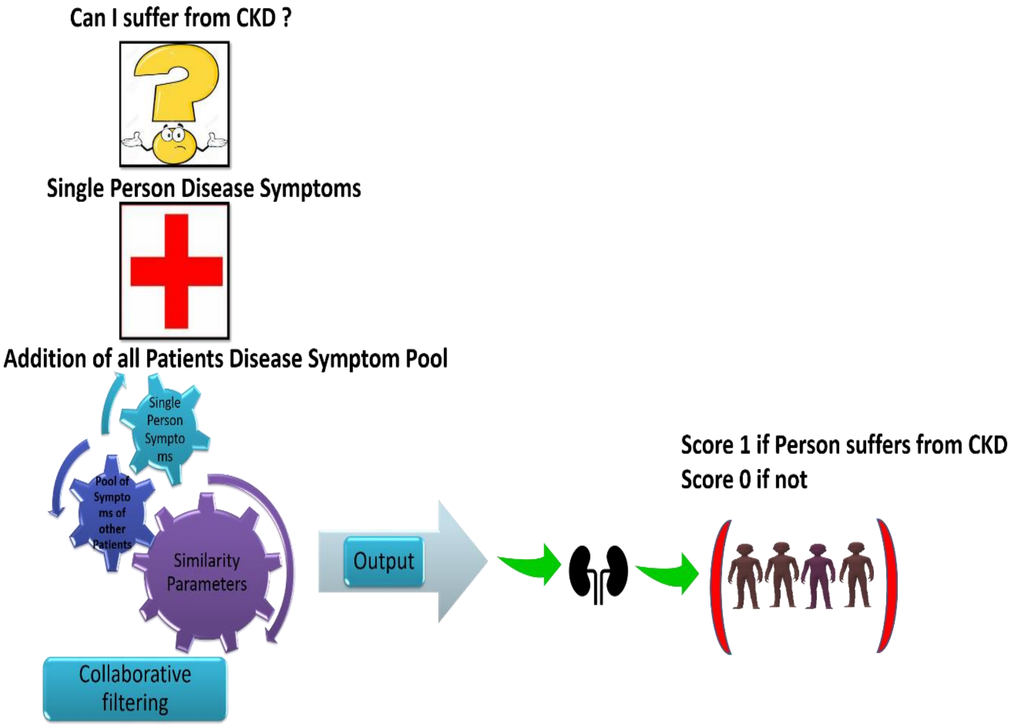

Can I suffer from CKD ?
Single Person Disease Symptoms
Addition of all Patients Disease Symptom Pool
Single Person Symptoms
Pool of Symptoms of other Patients
Similarity Parameters
Collaborative filtering
Output
Score 1 if Person suffers from CKD
Score 0 if not

Figure 4.5: Algorithm used to Extract the Symptoms of Similar Patients Based on Scores

Complexity Analysis and Methodology for Capacity Ratio

As presented in MySymptom pseudo-code in Fig. 4.6, the first step is a simple statement that takes the symptoms set of the patient, thus it requires only one unit of the time interval for its execution. The next step performs collaborative filtering to calculate the score of the patient. For the score calculation, this step compares the symptoms of the patient with the patients' data set, thus it takes units of the time interval or requires a number of comparisons. In the final step, the pseudo-code outputs the generated score of the patient therefore it takes a constant amount of time as it is a simple statement which only displays the output. Therefore, the overall complexity of the algorithm is calculated as O(n) because this algorithm requires comparisons for calculating the score.

Table 4.1: Representation of Scores Given to Patient Based on Similar Symptoms

No	Attributes	Patient 1	Patient 2
1	Age	53	20
2	Bp	90	70
3	Sg	1.02	1.02
4	Al	2	0
5	Su	0	0
6	RBC	Abnormal	Normal
7	PC	Abnormal	Normal
8	PCC	Present	Not present
9	BA	Not Present	Not Present
10	BGR	70	123
11	BU	107	44
12	SC	7.2	1
13	SOD	114	135
14	POT	3.7	3.8
15	HEMO	9.5	14.6
16	PCV	29	44
17	WC	12100	5500
18	RC	3.7	4.8
19	HTN	Yes	No
20	DM	Yes	No
21	CAD	No	No
22	Appet	Poor	Good
23	Pe	No	No
24	ANE	YES	No
25	Score	0.998	0.00003
26	Class	CKD	Not_CKD

Furthermore, the calculation of CR for all the three phases is performed on the Hadoop cluster. To estimate the CR of each DataNode for a particular job type is calculated after investigating the anatomy of Hadoop's environment to process a job. The profiling method used in this study maintains a separate ratio table at the master node that stores the information about every DataNode. After the execution of every job the information regarding input data size, execution time, job type performed on each DataNode is stored in the table at NameNode. The information of new DataNodes and their processing capacity is also updated. There are three situations for calculating the CR of each DataNode in a cluster as explained in Table 4.2.

1. The first time NameNode is not having any log for any DataNode.

2. On the cluster different jobs are executed and NameNode maintains each jobs information.

3. The job type is already performed on the cluster but a new DataNode is admitted.

```
$MySymptom\ Algorithm\ Pseudocode$
Input: $Individual\ Patient\ x\ having\ Symptoms\ \{S_1, S_2, \dots S_n\}$
Output: $Generate\ Individual\ Patient\ score$
            $where\ if\ score\ \cong 1\ means\ patient\ may\ have\ CKD;$
                $if\ score\ \cong 0\ means\ patient\ maynot\ have\ CKD$
1. $Input\ Patient\ x\ details, S \leftarrow \{S_1, S_2, \dots S_n\}$
2. $Apply\ Collaborative\ filtering, Collaborative_{output} \leftarrow (S, Dataset)$
3. $Output\ Patient_{score}$
```

Figure 4.6: P*seudocode* of MySymptom Algorithm

Table 4.2: Explanation of All Rounds Based on Jobs Distribution Performed in DataNodes by NameNodes CR Calculation After Applying Hadoop Distribution Algorithms

Rounds	**Processing step**	**Distribution**	**Master node information**	**Figure No.**
First-round	First time processing a job type, no information present of any DataNode	Equal distribution between all DataNodes.	NameNode maintains the CR for every DataNode.	Fig.4. 7 (A)
Second round	New job type arrives for processing	Equal distribution between all DataNodes for the new job type.	NameNode maintains the CR for every DataNode.	Fig. 4.7 (B)
Third round	New DataNode added in the cluster and performing previous job type.	Minimum jobs are given to the new DataNode for this round and other DataNode gets based on CR.	NameNode updates the CR	Fig. 4.7 (C)

To estimate the CR of every DataNode MapReduce job is performed for the first time on each DataNode. As shown in Fig. 4.7 (A) DataNode 2 and 3 are 1.47 times faster than DataNode 1. The CR of DataNode 1, 2, 3 is 0.56, 0.7, 0.7, respectively estimated based on algorithm 4.2. The next round is performed for a different job type MySymptom task on the same cluster resources. Figure 4.7 (B) shows that for this job type again DataNode 1 is the slowest performer of the cluster. DataNode 2 and 3 are 1.25 times faster than DataNode 1. As the CR of these three DataNodes 1, 2, and 3 is 0.611, 0.88, and 0.88, respectively. In the next round again MapReduce job is performed but this time the cluster resources are changed as there is a new DataNode in the cluster. For this round, the newly admitted node (DataNode 4) and DataNode 1 get minimum jobs to perform. Figure. 4.7 (C) depicted that DataNode 4 is the fastest performer of a cluster in comparison to the existing DataNodes.

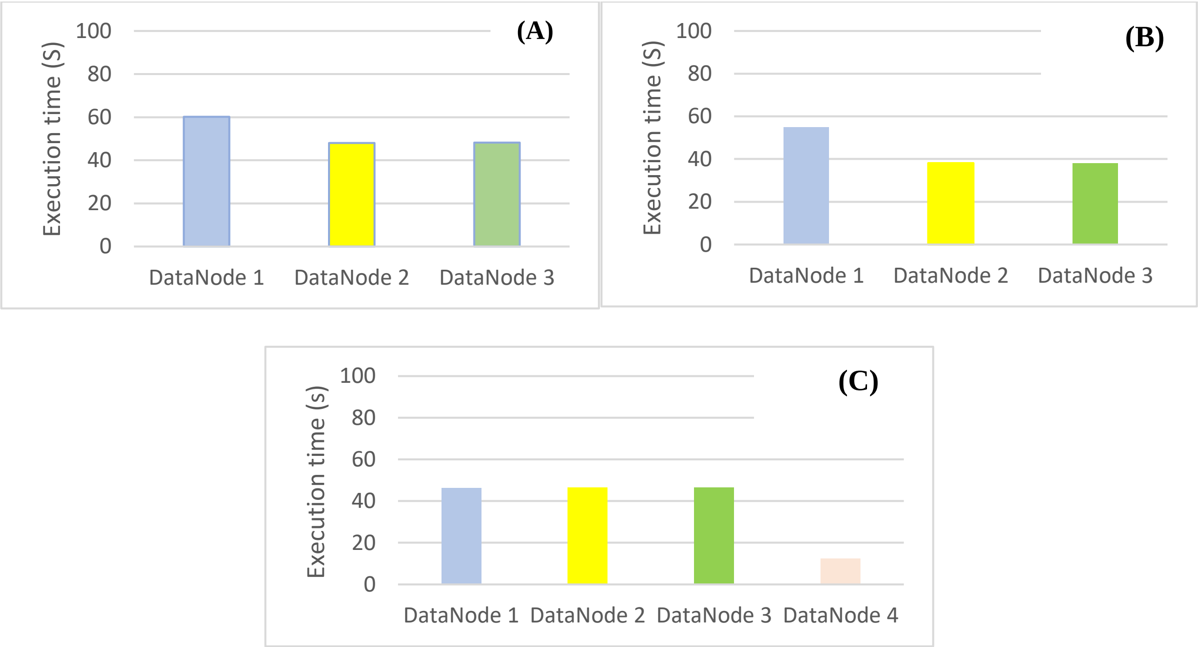


Figure 4.7: Execution Time of DataNodes for: (A) The First Round, (B) Second Round Showing the Results for the New Job Type And (C) Third Round Showing Performance of the Hadoop Cluster After Horizontal Scaling

4.2.1.5 Data Analysis Layer (DAL)

To develop an intelligent healthcare system ML algorithm plays an important role. Different ML algorithms woks on various logical attributes to develop a good predictive system. In this section, we discuss the use of various ML algorithms and compare their results to get the best algorithm out of them. Figure 4.8 is visualizing the working of the proposed model that how it takes data from a centralized disease pool and then distributes it to get a final aggregated result.

4.2.1.6 Report Generation Layer (RGL)

In the RGL different reports are prepared for doctors, patients, government officials so that they can understand all pros and cons of their problem. The best representation of the report is the graphical and tabular format. The tool used for analysis is generally present the result in both ways. This feature saves the time of report analysis for understanding the report without putting much effort.

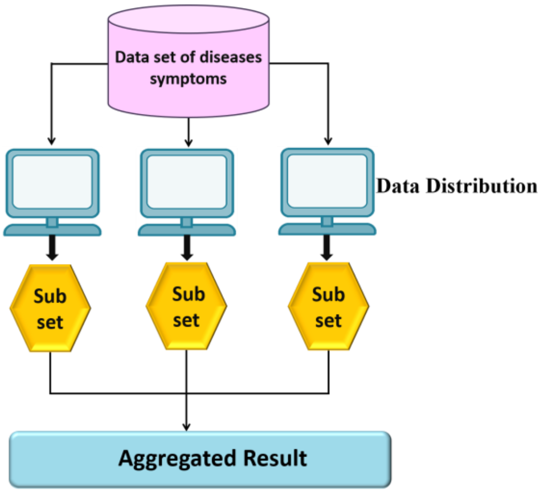


Figure 4.8: Working of Distributed Model for Faster Data Processing

4.3 Data Analysis

4.3.1 Dataset and Attributes

To show this correlation of attributes we are using a UCI [235] and data.CMS.gov dataset of CKD disease. There are 24 parameters as shown in Table 4.3 that can cause CKD and they all are used to create a predictive model. The dataset contains 25,190 individuals, among them, 25,040 suffer from CKD. Kidney disease having total five stages and these stages are based on the glomerular filtration rate (GFR). GFR is used to measure the status of a patient's kidney and it tells the doctors about the kidney function of a patient.

- **Urea in Blood:** Urea is a chemical waste product and it generates when nitrogen is mixed with elements like carbon, oxygen, and hydrogen. Urea travels from the liver to the kidney through the bloodstream. The function of a healthy kidney is to filter out the urea and other waste products from the patient's body. All the waste products filter out through urine. Urea level is tested by the blood urea nitrogen (BUN) test which tells the urea nitrogen levels are higher or not.

Table 4.3: Parameters Responsible for Chronic Kidney Disease

S.No.	Attributes abbreviation	Attributes full form	Data Type	Attribute Information
1	Age	Age	Numerical	Age in years
2	BP	Blood Pressure, mm/HG	Numerical	Mm/Hg
3	Sg	Specific Gravity	Nominal	1.005,1.010,1.015,1.020,1.025
4	Al	Albumin	Nominal	0,1,2,3,4,5
5	SU	Sugar	Nominal	0,1,2,3,4,5
6	Rbc	Red Blood Cells	Nominal	Normal, abnormal
7	Pc	Pus cell	Nominal	Normal, abnormal
8	Pcc	Pus cell Clumps	Nominal	Present, not present
9	Ba	Bacteria	Nominal	Present, not present
10	Bgr	Blood Glucose Random	Numerical	mgs/dl
11	Bu	Blood Urea	Numerical	mgs/dl
12	Sc	Serum Creatinine	Numerical	mgs/dl
13	Sod	Sodium	Numerical	mEq/L
14	Pot	Potassium	Numerical	mEq/L
15	Hemo	Hemoglobin	Numerical	Gms
16	Pcv	Packed Cell Volume	Numerical	-
17	Wc	White Blood Cell Count	Numerical	cell/cumm
18	Rc	Red blood Cell Count	Numerical	millions/cmm
19	Htn	Hypertension	Nominal	Yes, No
20	Dm	Diabetes Mellitus	Nominal	Yes, No
21	Cad	Coronary Artery Disease	Nominal	Yes, No
22	Apt	Appetite	Nominal	Good, Poor
23	Pe	Pedal Edema	Nominal	Yes, No
24	Ane	Anemia	Nominal	Yes, No
25	Class	Class	Nominal	CKD, not CKD

- **Hematuria:** Hematuria is a stage in which red blood cells are present in urine. Generally, urine does not contain red blood cells they are being prevented by the kidney from entering into the urine. Hematuria happens when a kidney or other parts of the urinary tract does not perform well. The general causes of hematuria are mineral imbalances in the urine, inherited diseases like Polycystic Kidney Disease (PKD), inflammation of the kidney.

- **WBC:** The higher rate of white blood cells or (pus cells) in the urine can cause a kidney infection.
- **Polycystic Kidney Disease (PKD):** PKD refers to an inherited disorder and in this disease, a large number of cysts grow in the kidney. These cysts can cause the kidney to get enlarge in size and due to these kidney losses its function over time. A cyst contains fluid in them, and its size may vary. If someone has a large number of cysts, then it can also damage the kidney. Generally, PKD is a result of an abnormal dominant gene of an affected parent to its children.

4.4 Case Study: The Integrative Investigation of CKD

In a country like India spreading awareness for organ donation is a big task. Approximately 70% of its population lives in rural areas according to the census. In India Transplantation of Human Organ Act, was passed in 1994. This Act mainly focuses on two points (i) to regulate the storage, removal, and transplantation of human organs for therapeutic use [236]. (ii) For preventing commercial dealing of human organs. There is a huge need for spreading awareness in India. Around 2,20,000 people are demanding for transplantation of kidney where only 15,000 people were getting their kidney transplanted [237, 238]. Kidney transplantation is required to provide a better life for patients who suffer from end-stage renal diseases (ESRD). In developing countries, the renal transplantation rates are considerably low than in developed countries. According to reports [239] it can be easily visualized in Fig. 4.9 that there is an urgent requirement of organ donation awareness in India to stop illegal organ trading. The reason for this could be lack of awareness, low education level, no clear national policy, dearth of an organized system for organ retrieval from deceased donors, and deficiency of functional dialysis and transplant unit. The government should take the right action to overcome these drawbacks. Our proposed model helps them to implement the important points in action for CKD, ESRD, and transplantation registries. Kidney transplantation is not the only issue that should be addressed but after the transplantation follow-up is very important for chronically ill patients. A powerful tool can help to provide superior patient care and also help them to detect the disease in advance. These types of tools strengthen the economic background of the country by reducing the additional expenditure on chronic diseases like a kidney.

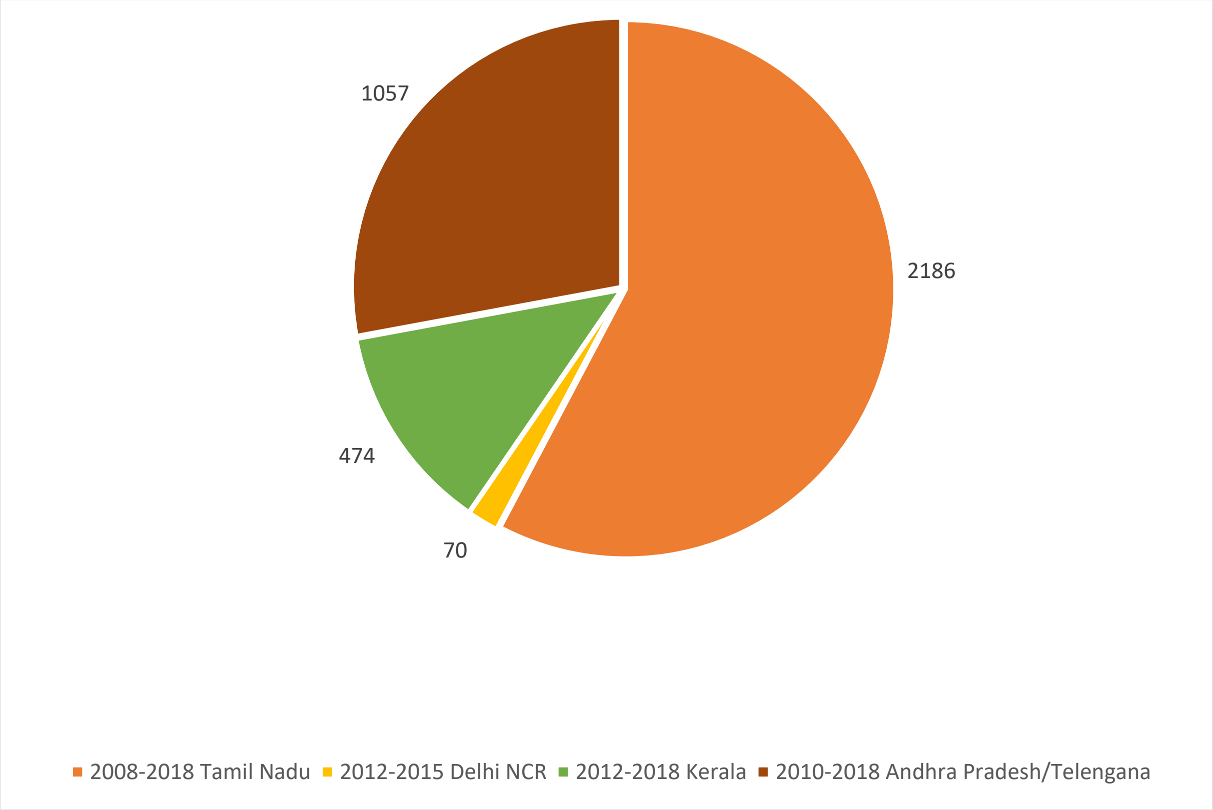


Figure 4.9: Statistical Distribution Representing Kidney Donation in India

4.5 Experimental Setup

This section explains the experimental setup of HCMP architecture to predict CKD as summarized in Table 4.4. HCMP architecture works on six layers and every layer is dedicated to perform a particular task.

Table 4.4: Techniques and Methods Used at Different Layers of HCMP Architecture

Tools/Technique used	**Version**	**Objective/Work**	**Working layer of architecture**
Hadoop	2.6.4	Data storage and data management	Layer 2 and 3
MySymptom algorithm	-	Data filtering	Layer 4
RapidMiner	8.0	Data analysis	Layer 5
RapidMiner	8.0	Report generation	Layer 6

Hadoop is used for data placement and data management in layers two and three. The experimental environment of the Hadoop cluster is real where four DataNode and one NameNode were used. The NameNode is the master node of the cluster and the four DataNode are the slave node of the cluster. The configuration of each node is shown in Table 4.5. The version of Hadoop used for all nodes of the cluster is 2.6.4 and the operating system is Ubuntu 14.04. The internal memory and the disk space vary for all nodes of the cluster. At the same time, MySymptom algorithm works for data processing in layer 4. The work of the algorithm is to take the data from the central repository and check the similarity between the patients and give scores to them based on their similarities. Experimental results described in this study were obtained by using six standard ML algorithms provided by RapidMiner. Here, the Radoop extension is used to access the functionality of the Hadoop cluster. Hence, to use the Radoop extension initially we have to configure the connections between RapidMiner studio and RapidMiner Radoop. After this RapidMiner Radoop needs to get connected to an already installed Hadoop cluster by giving some networking permissions. Through these permissions, the RapidMiner Radoop client gets access to the Hadoop cluster and can be managed through the Hadoop design view or data view.

Table 4.5: Specification of DataNodes and The Experimental Environment of the Hadoop Cluster

Node	**CPU**	**Memory**	**Disk**	**Operating system**	**Hadoop version**	**Java**
Master node	Intel Core2duo	3.7 GB	43.7 GB	Ubuntu 14.04	2.6.4	JDK 7
Slave node 1	Intel Core2duo	1.9 GB	56.4 GB	Ubuntu 14.04	2.6.4	JDK 7
Slave node 2	Intel Core2duo	3.7 GB	99.0 GB	Ubuntu 14.04	2.6.4	JDK 7
Slave node 3	Intel Core2duo	3.7 GB	45.3 GB	Ubuntu 14.04	2.6.4	JDK 7
Slave node 4	Intel Core I5	8.0 GB	499 GB	Ubuntu 14.04	2.6.4	JDK 7

4.6 Results and Discussion

Data analysis is conducted by using ML algorithms to predict CKD. Six different classification algorithms namely DT, KNN classification, Kernel Distribution (Naïve Bayes), Simple Distribution (Naive Bayes), RF, and RT are used. The performance of different classifiers was measured as: recall, precision, accuracy, AUC optimistic, Area Under the Curve (AUC), and

AUC pessimistic. The calculation of all the above-mentioned performance measures has been done by a confusion matrix [240]. Where TP, TN, FP, and FN means, true positive (correctly predicting the number of cases as required), true negative (correctly predicting the negative number of cases), false positive (incorrectly predicted the negative case as positive), false negative (incorrectly predicted the positive case as negative), respectively. The mathematical relation to calculate accuracy, precision, recall/sensitivity, specificity, and false positive rate (FPR) is explained in Equation (4.1) to (4.6).

$$\text{Accuracy}= \frac{TP+TN}{TP+FP+TN+FN} \qquad (4.1)$$

$$\text{Precision}=\frac{TP}{TP+FP} \qquad (4.2)$$

$$\text{Sensitivity/Recall/True Positive Rate (TPR)}= \frac{TP}{\text{TP+FN}} \qquad (4.3)$$

$$\text{Specificity/True Negative Rate}= \frac{TN}{TN+FP} \qquad (4.4)$$

$$\text{FPR}=\frac{FP}{TN+FP} = 1 - Specificity \qquad (4.5)$$

$$\text{AUC}=\hat{A} = \frac{S_0 - n_0(n_0+1)/2}{n_0 n_1} \qquad (4.6)$$

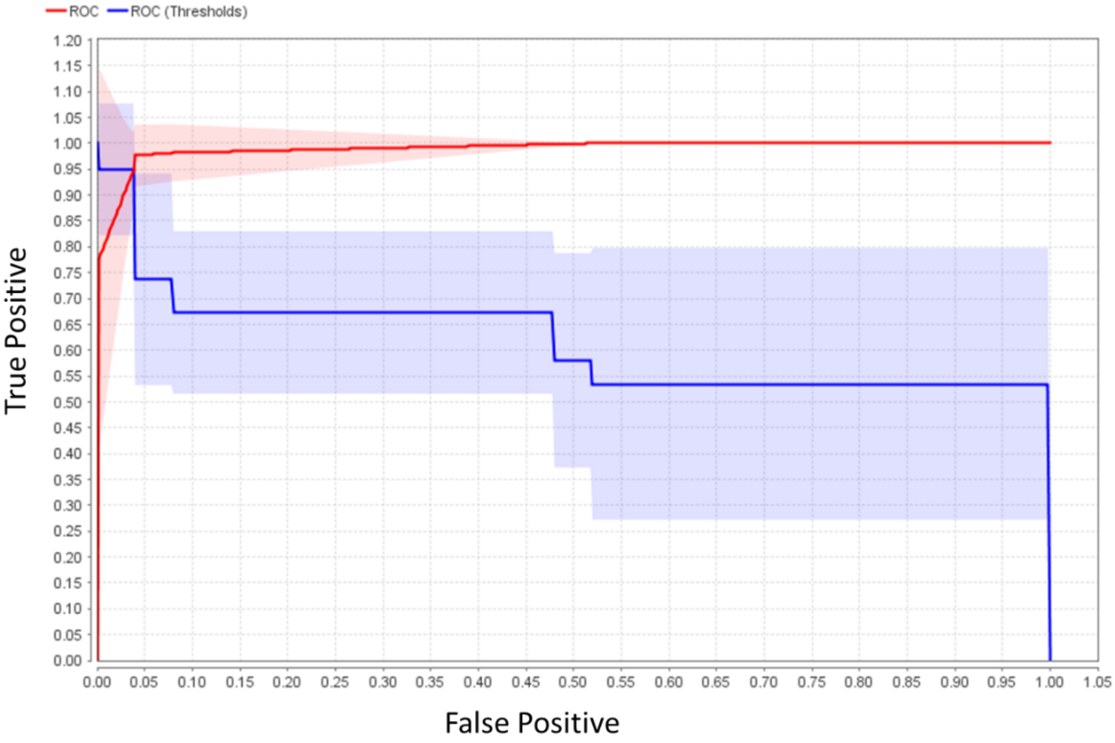
4.6.1 Analysis Based on AUC-ROC Plot

The Receiver Operator Characteristics (ROC) graph is used for evaluation of classification problems. ROC is a probability curve that plots TPR against FPR at different threshold values. Where AUC tells about the ability of a classifier and checks whether a classifier can distinguish between the classes perfectly or not. Higher AUC tells that the model performance is best in differentiating between the negative and positive classes. In Equation 4.6, n_0 and n_1 represents the number of positive values and negative values [241]. Where $S_0 = \sum r_i$ and r_i is the position of the ith positive value in the list. AUC curve gives a better result if the area is closer to 1. According to Ling et al. and Bradley et al. [242], AUC is a better evaluation method for classifiers than accuracy as it is found that it has increased sensitivity in ANOVA (analysis of variance) test [243, 244]. The results of different performance measures calculated by various algorithms are shown in Table 4.6. The AUC for DT, KNN Classification, Kernel Distribution (Naïve Bayes), Simple Distribution (Naïve Bayes), RF and RT are shown in Fig. 4.10 (A-F), respectively. Figure 4.10 (A) shows the AUC of DT where it can be seen that the difference between sensitivity and specificity is very less. Here, AUC of KNN revels that the classifier is able to score a good sensitivity score in comparison to the specificity score in Fig. 4.10(B). The graph in Fig. 4.10 (C & D) for Kernel Distribution (Naïve Bayes) and Simple Distribution (Naive Bayes) shows AUC that the perfect result for sensitivity and specificity is observed. It means models can classify all positive cases of CKD as positive and all negative cases as negative. Because of that the sensitivity also becomes high for these models. In the case of RF and RT shown in Fig. 4.10 (E & F), the sensitivity values are almost the same but the specificity is good for RF in comparison to RT. It means the RT model is not able to do the better classification of false-positive cases. The confusion matrix for all the classifiers has been given in Table 4.7.

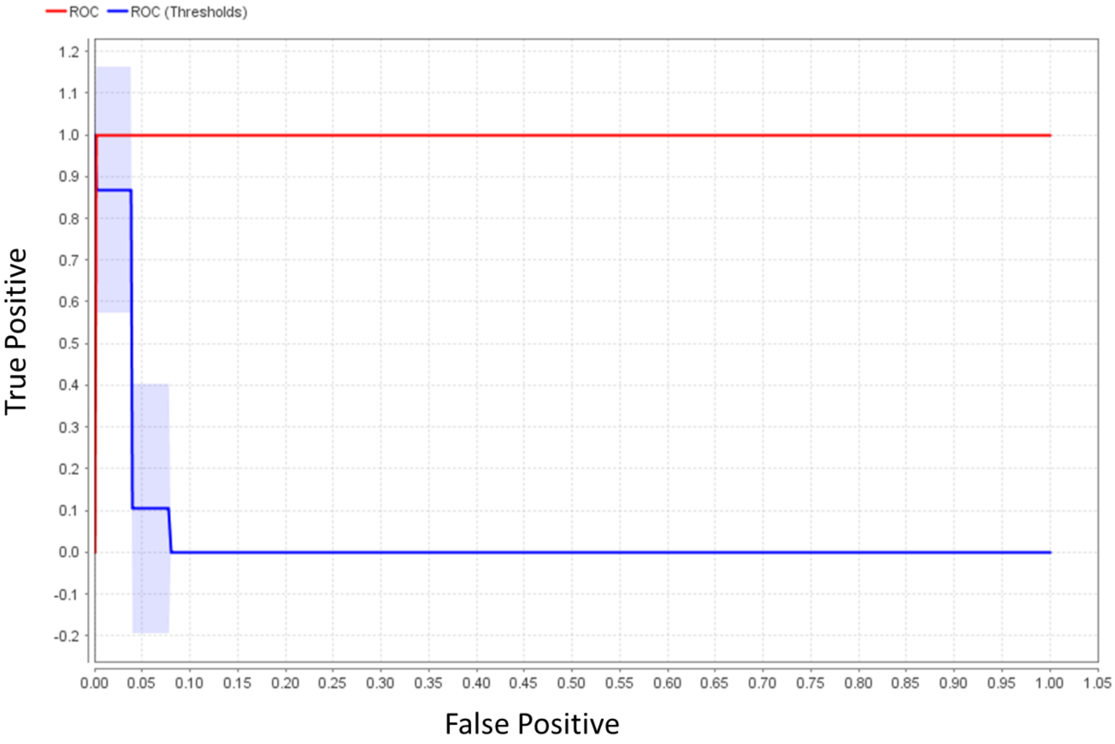
Table 4.6: Comparison of Various Prediction Algorithms for Chronic Kidney Disease

Classifier	**Accuracy**	**Precision**	**Recall**	**AUC OPTIMISTIC**	**AUC**	**AUC PESSIMISCTIC**
DECISION TREE	97.25%	95.15%	98.00%	0.998 ± 0.003	0.990 ± 0.017	0.981 ± 0.032
KNN-CLASSIFICATION	97.50%	96.32%	97.33%	0.999 ± 0.002	0.500 ± 0.000	0.950 ± 0.067
Kernel Distribution (Naïve Bayes)	99.50%	99.38%	99.33%	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Simple Distribution (Naive Bayes)	99.25%	99.38%	98.67%	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Random Forest	97.25%	99.33%	93.33%	0.994 ± 0.006	0.993 ± 0.007	0.993 ± 0.008
Random Tree	77.25%	67.05%	77.33%	0.988 ± 0.011	0.787 ± 0.163	0.587 ± 0.334

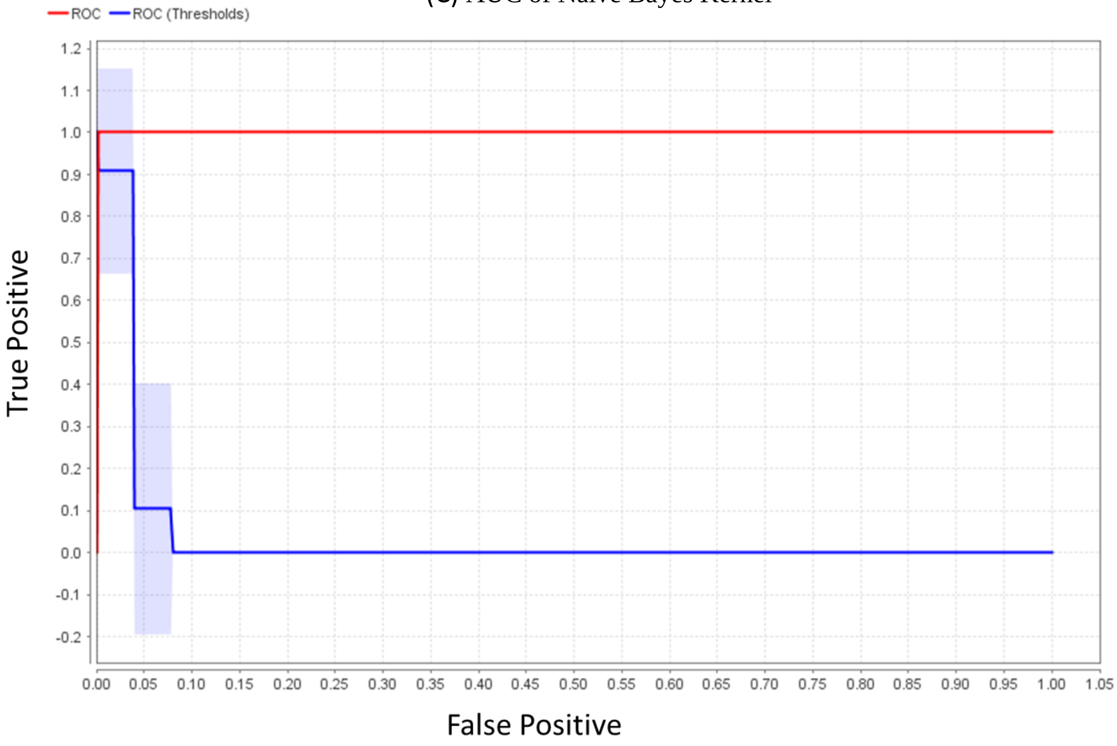
(A) AUC of Decision Tree



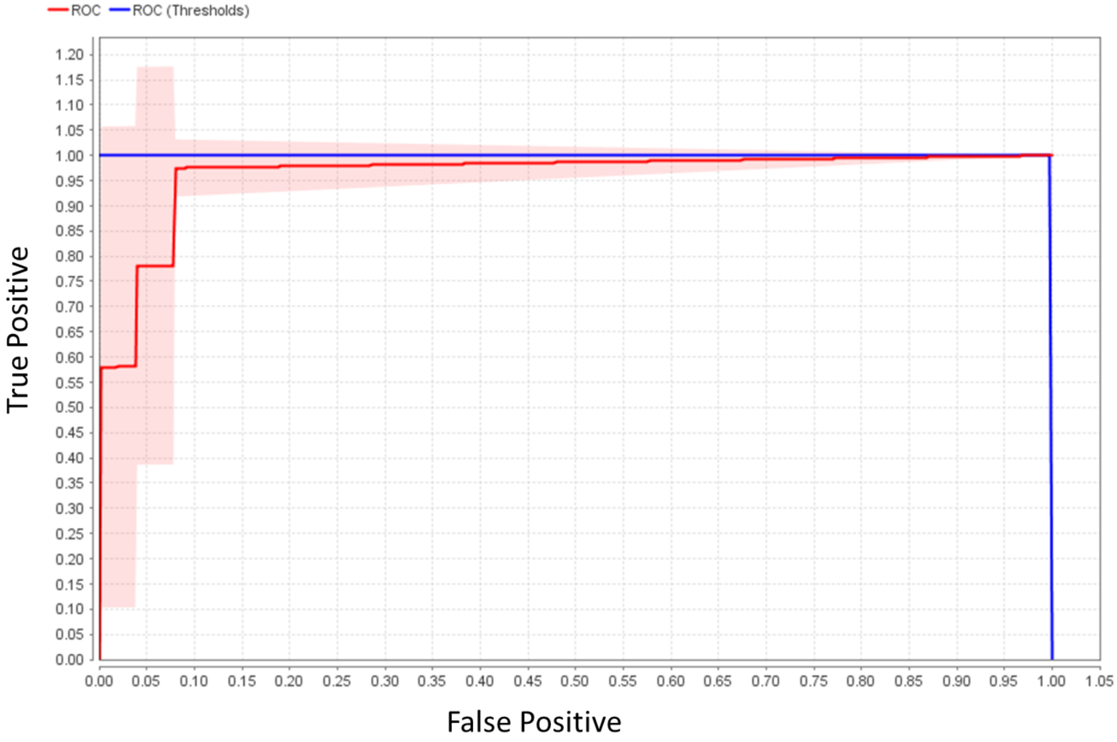
(B) AUC of Simple Distribution Naïve Bayes



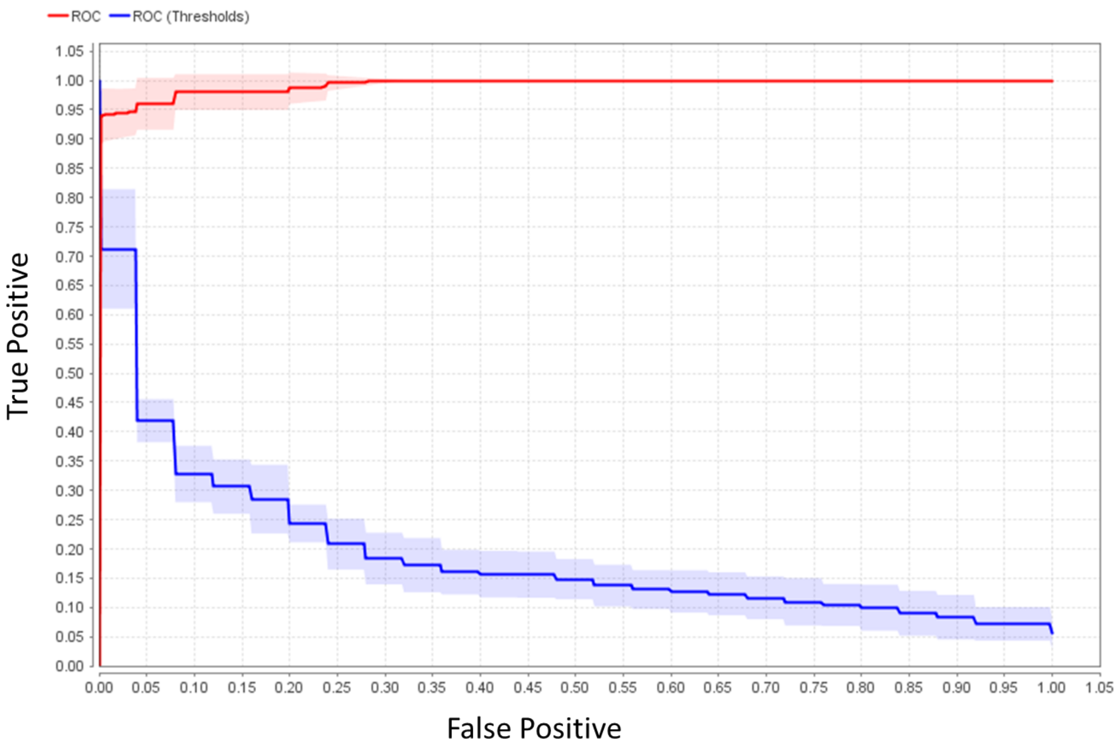
(C) AUC of Naïve Bayes Kernel



(D) AUC of KNN



(E) AUC of Random Forest



(F) AUC of Random Tree.

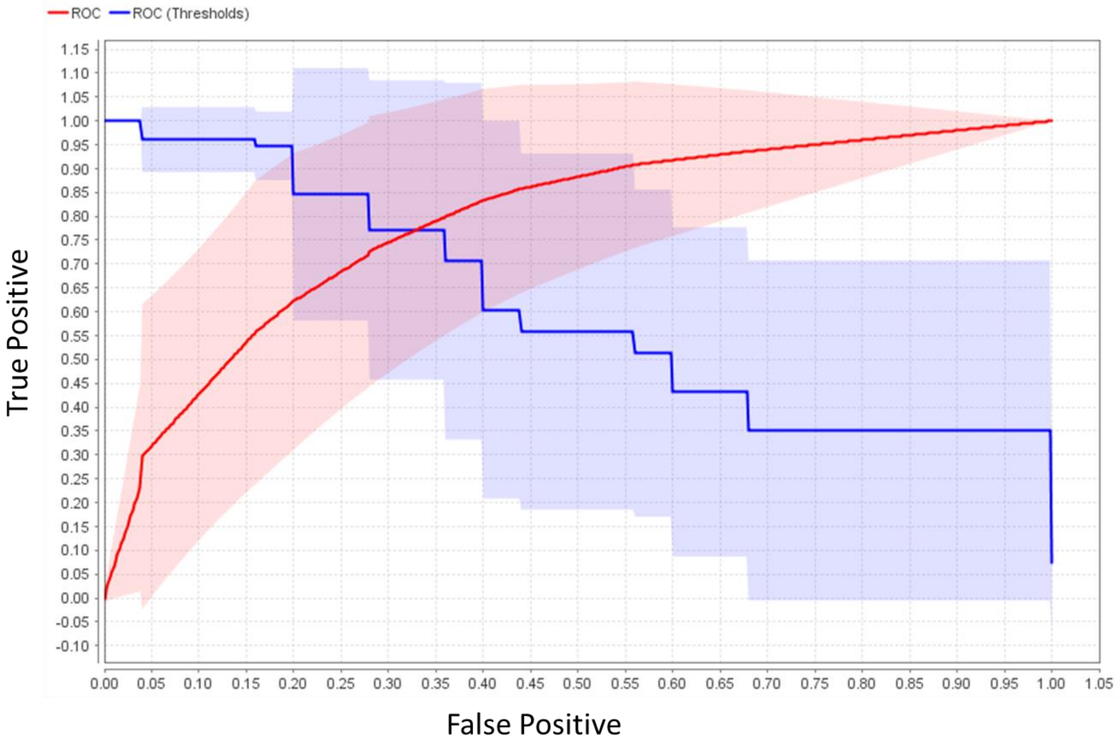


Figure 4.10: (A) AUC of Decision Tree (B) AUC of Simple Distribution Naïve Bayes (C) AUC of Naïve Bayes Kernel (D) AUC of KNN (E) AUC of Random forest (F) AUC of Random Tree

As from the extracted results reported in Table 4.6, the Kernel Distribution (Naïve Bayes) produces better results with an accuracy of 99.5%, precision of 99.38%, recall 99.33%, and AUC as perfect 1. The Kernel Distribution (Naïve Bayes) and Simple Distribution (Naive Bayes) produces better results in comparison to other models may be due to the non-overlapping and independent features property of NB. Whereas Naïve Bayes (Kernel) is used for numerical attributes. A kernel is a weighting function that works for nonparametric estimation methods and it is used for estimation of random variable density function in kernel density estimation.

Table 4.7: Confusion Matrix of Different Classifiers

Classifier	Confusion matrix	
DECISION TREE	25032	3
	8	147
KNN-CLASSIFICATION	25034	4
	6	146
Kernel Distribution (Naïve Bayes)	25039	1
	1	149
Simple Distribution (Naive Bayes)	25039	2
	1	148
Random Forest	25039	10
	1	140
Random Tree	24983	34
	57	116

The working of NB classifier [245, 246] is based on Bayes theorem. Bayes theorem is useful when works with conditional probability.

$$\mathrm{P(A/B)} = \frac{\mathrm{P(B/A)xP(A)}}{\mathrm{P(B)}} \qquad (4.7)$$

After putting the Naïve assumption to the Bayes theorem that is independent among the features used to split evidence into the independent parts. The two independent events are A and B.

$$\mathrm{P\ (A, B) = P(A)P(B)} \qquad (4.8)$$

Then,$P(y|x_1, \dots\dots, x_n) = \frac{\mathrm{P}(x_1|y)P(x_2/y)\dots\dots P(x_n|y)P(y)}{P(x_1)P(x_2)\dots\dots P(x_n)}$ (4.9)

It can be written as $\mathrm{P(y|x1, \ldots\ldots, xn)} = \frac{\mathrm{P(y)}\pi^{n}i=1P(xi|y)}{\mathrm{P(x1)P(x2)\ldots.P(xn)}}$ (4.10)

In above Equation, for a given input we can remove denominator, If the denominator remains constant and Equation 4.10 can be represented as;

$P(y|x_1, \ldots\ldots, x_n) \propto P(y)\pi^{n}i=1P(x_i)|y$ (4.11)

Hence, for making a classifier model, the probability is given for all set of inputs of the class variable y has been calculated which takes the maximum probability as output.

$Y=argmaxy\ P(y)\pi^{n}$ i=1 P(x_i|y) (4.12)

The last step is to calculate the class probability, *P(y)* and conditional probability P(x_i|y). One of the advantages of the NB classifier is that a small amount of training data is required for estimating variances and means of the variables required for classification. One more property of NB is that conditional independence assumption helps to reduce the complexity of NB and which makes it a robust model. The conditional independence assumption of NB helps it to overcome the overfitting problem. Ling et al. [241] has explained the performance of NB by comparing the working of it with some state-of-the-art algorithms. The authors have performed experiments on 36 different datasets from UCI and has calculated the average AUC for them. The results produced by their study show that NB attains the highest AUC average over all other algorithms. The above advantage of NB also justified in our experimental results where the AUC of Kernel Distribution (Naïve Bayes) and Simple Distribution (Naive Bayes) is perfect 1 in comparison to other algorithms.

4.6.2 Comparison with Other Proposed Techniques

On comparing various studies [113, 110, 116, 111, 118, 237, 119], the results of these studies have indicated that ML algorithms and different architecture help to improve the healthcare sector. According to the performance measures such as accuracy, recall, precision and AUC, the results generated in the present study are improved. Furthermore, the architecture used in the present study uses Hadoop for data storage and data management by applying an improved data placement strategy for better healthcare services.

In a recent study by Karthikayan et al. (2019) [113] a multi-disease prediction model has been proposed. They have used the SVM-Radial bias kernel method and compared it with SVM-Polynomial, SVM-Linear, DT, and RF methods. The authors have experimented using three classification data sets of three different diseases such as CKD, Diabetes, and Heart Disease.

Their experimental results have shown that the SVM-Radial bias kernel method is the best approach among other classification algorithms. Our proposed architecture HCMP works on different layers to provide a better healthcare system using a parallel processing environment to improve the data processing rate. The results of the data analysis layer of our architecture are compared with the results of the data processing layer of Karthikayan et al. [113] in terms of performance measurement parameters: accuracy, recall, precision and AUC are shown in Table 4.8. Hence, the value of accuracy, precision and AUC shown improved except for recall in the present study.

In another study of Ahmed Abdelaziz et.al. (2019) [110], the ML-based model is used to improve healthcare services (HCS) in a cloud environment. This model works for HCS on a cloud to optimize the virtual machines (VMs) selection using parallel particle swarm optimization. In addition to this, they have created a new model for CKD diagnosis to measure the performance of the virtual machines model. Hence, the VMs selection model increases the data processing rate for HCS. In the present study, we have also improved the HCS with the help of improved data placement algorithms for Hadoop. The comparison of accuracy, precision, recall, and AUC of Ahmed Abdelaziz et.al. study with the present study is shown in Table 4.8. The result shows an improvement in the values of accuracy, precision, and AUC. However, the Recall value of [110] is better than the proposed study.

The study of Pramila Arulanthu et. al. (2020) [116] has proposed an intelligent medical decision support system using IoT & cloud for CKD. The model presented by them involves various stages like data gathering, data processing, and data classification. The LR model was used for CKD classification after tuning its parameters. However, the result of the proposed algorithm of Pramila Arulanthu et.al. is compared with other classification algorithms and shows improvement in accuracy. Our study also uses different layers of architecture like data collection, storage of data, management of data, data processing, and data analysis. For data analysis of CKD Kernel Distribution (Naïve Bayes) produces the best results. On comparing the accuracy, precision, recall and AUC values of the classification model used by Pramila Arulanthu et.al. [116], the present model produces the best results as shown in Table 4.8.

Moreover, in the study of Mohamed Elhoseny et al. [111] a novel approach for intelligent diagnostic prediction of CKD using density-based feature selection algorithm with Ant colony-based optimization algorithm. The comparison based on accuracy, precision, recall and AUC values is depicted in Table 4.8 and it indicates the improvement in the result than the D-ACO approach. In another study of Yafeng Ren et al. (2019) [237] authors propose a hybrid neural network using Auto encoder and bidirectional long short-term memory (BiLSTM) networks that captures the textual and numerical information in EHR. The raw EHR data has been used by them that belongs to hypertension patients for knowing the risk of CKD in those patients. The results generated in their study attains 0.87 accuracy, 0.91 recall, 0.93 precision, and 0.895 AUC. Hence, the accuracy, recall, precision and AUC values of our proposed study is much better than the accuracy of Yafeng Ren et al.

The study presented by Vásquez-Morales et al. (2019) [118] used a neural network-based model for predicting the risk of developing CKD in a person. The authors used different ML algorithms and calculate performance metrics and the best performance metrics are produced by NN. At the same time in another study by Kriplani et al. (2019) [119] author uses a deep neural network (DNN) based prediction method for CKD prediction. They use six classification algorithms and among them DNN gives the best result. Therefore, an improved value of accuracy, precision, recall and AUC are observed in our proposed study as shown in Table 4.8. Figure 4.11 shows the comparison of accuracy, recall, precision and AUC of present study with the other studies (classification techniques).

Table 4.8: Comparison of Classification Accuracy, Recall, Precision and AUC

Dataset	Accuracy	Recall	Precision	AUC
CKD	0.995 (proposed)	0.993 (proposed)	0.993 (proposed)	1 (proposed)
	0.973 [113]	1 [113]	0.96 [113]	0.959[113]
	0.97 [110]	1 [110]	0.97 [110]	0.997 [110]
	0.98 [116]	0.97 [116]	0.95[116]	0.967[116]
	0.97[111]	0.98 [111]	0.96[111]	0.95[111]
	0.972 [119]	1 [119]	0.95 [119]	0.969[119]
	0.87[237]	0.91[237]	0.93[237]	0.895[237]
	0.965[118]	0.975 [118]	0.95[118]	0.98[118]

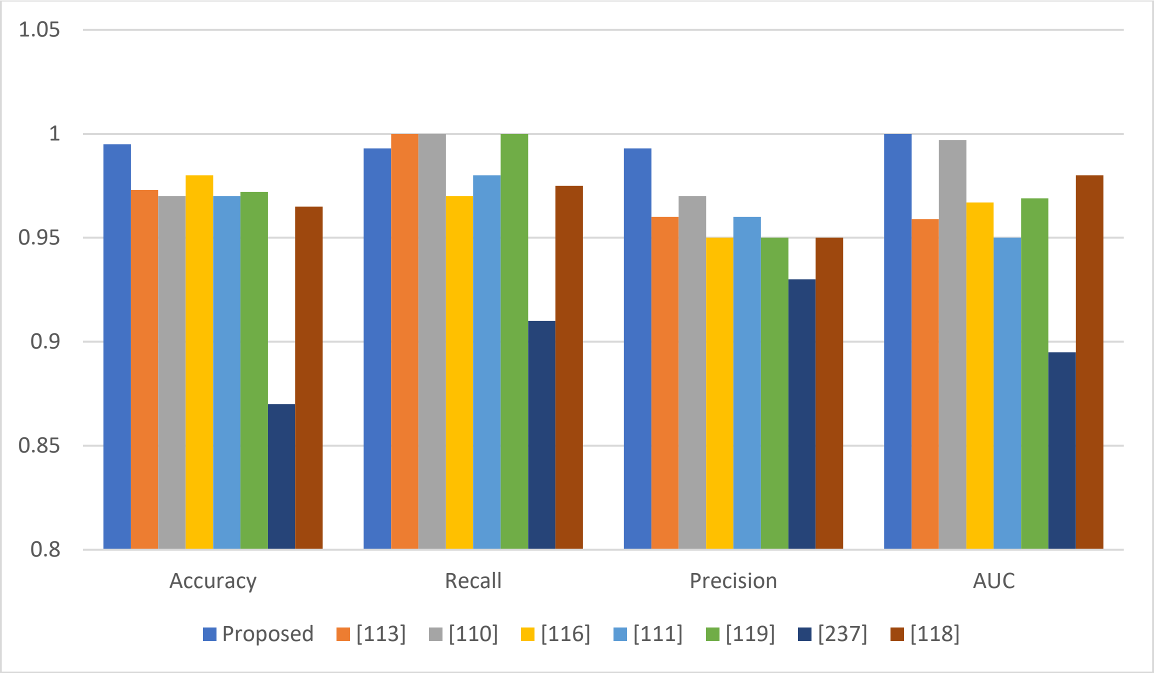


Figure 4.11: Comparison of Performance Measures with Other Models

4.7 Summary

A new HCMP architecture has been successfully devised to predict CKD with improved healthcare services. This HCMP architecture has worked on six layers. The data storage and management layer was performed on heterogeneous Hadoop cluster. Also, the CR of each cluster's DataNode for a particular job type was estimated after investigating the anatomy of Hadoop's default architecture. The profiling methods were used to consider three situations for calculating the CR of each DataNode in the cluster such as the NameNode without having any information of DataNode's computing capacity, running the type of job first time on the cluster, and the stratergy of NameNode to handle the case of horizontal scaling.

Hence, firstly the MapReduce task was performed for data processing keeping equal distribution on each DataNodes. Secondly, the MySymptom task has been performed for collaborative filtering of patients' data by equal distribution on each DataNodes. In the third round, a new DataNode was added and a MapReduce job was performed after distributing the data based on CR. The results have revealed that DataNode 2 and DataNode 3 are 1.47 times faster than DataNode 1 for the MapReduce job and 1.25 times faster for the MySymptom job. Here, it was observed that the new admitted DataNode 4 was the fastest among all DataNodes of the heterogeneous Hadoop cluster. After the data processing layer, the prediction of CKD was performed by the data analysis layer and finally, it was ended at the report generation layer. Therefore, it has been noticed that among all the classification algorithms used in this study, the best results were obtained by Kernel Distributed (Naïve Bayes). On comparing with other studies by the various researchers, the results in this investigation revealed that the classification performance parameters such as accuracy, precision and AUC have been improved by 2.2%, 2.3% and 0.3%, respectively. However, the recall of other studies was 0.7% better than the present study.

Moreover, for multiple disease prediction the present proposed architecture has been further extended that will be able to handle multiple diseases along with the hybrid machine learning model at data analysis layer for better prediction.

CHAPTER 5

D-SMO Based Multi Disease Prediction Model

It has been observed from the literature review that there is a requirement for a healthcare architecture that performs multiple disease detection using a robust method for data placement and management. This chapter proposes a Hadoop-based Optimal Healthcare Classification Multi-Disease Diagnostic (OHCMDD) architecture for handling multiple diseases database with reduced feature set. The proposed architecture also handles the DataNodes decommissioning problem. In addition, it provides a complete heterogeneous distributed environment for multi-disease detection. The experimental results, performance and analysis evaluation shows that the proposed architecture provides a reliable and feasible environment for healthcare.

5.1 Introduction

In recent years, healthcare-related big data is growing exponentially and it creates a big challenge in front of the old conventional methods. Thus, BDA and health 4.0 approach opens the doors of opportunity to examine large datasets for finding the hidden patterns, values and uncover new insights [247]. The BDA techniques are used in different fields like healthcare, fraud detection, weather forecasting, and logistic delivery. At the same time, health 4.0 helps in the digitization of laboratories and allows automation in different processes used in hospitals and the general health sector. In healthcare, BDA and health 4.0 helps to deal with the problems associated with the management of exponentially generated medical data and extracting relevant information from it. The proper handling of data is very important in terms of data storage and data processing. This leverages the patients and healthcare practitioners to understand the complex part of the data in less time. Health 4.0 integrates the principles of Industry 4.0 (I4.0) and this helps the physicians to improve their efficiency by optimizing resources for providing expeditious improvement in the patients' health conditions [248].

The genesis of health 4.0 spreads its dawn in this world with the emergence of I4.0. I4.0 aims to bring the revolution by digitization in the manufacturing sector [249]. I4.0 also helps to connect the physical world with the virtual world in real-time. Virtualization is also happening in health due to the emergence of ICT. The use of different technologies in healthcare causes the generation of an abundance amount of health-related big data.

To achieve the target of providing better health care services, various technologies need to be concatenated. The conjunction of BDA, health 4.0, RAMI 4.0 [250], and ML make sure that all the involved stack holders can understand the requirements of each other. Figure 5.1 presents the association of the seven V's of big data and different technologies that help to achieve the goal of health 4.0.

To provide better healthcare services the orchestration of assorted participants like physicians, laboratories, pharmacies, EHR and medical journal articles needs to be done properly [251]. Figure 5.2 shows the use of different technologies used in health 4.0. This is a technological challenge to build a unified and smart architecture for healthcare. In this study, an OHCMDD architecture is created. This framework helps to manage the health-related data properly and is able to diagnose different diseases. The proposed architecture consists of four layers for fast and efficient analysis of diseases like CKD, diabetes, and heart disease. The hybrid of the metaheuristic D-SMO algorithm has been proposed to produce better classification results. The contribution of this article is two-fold that is (1) OHCMDD architecture is proposed to collaborate different technologies for providing improved healthcare services. (2) D-SMO algorithm is proposed for data analysis.

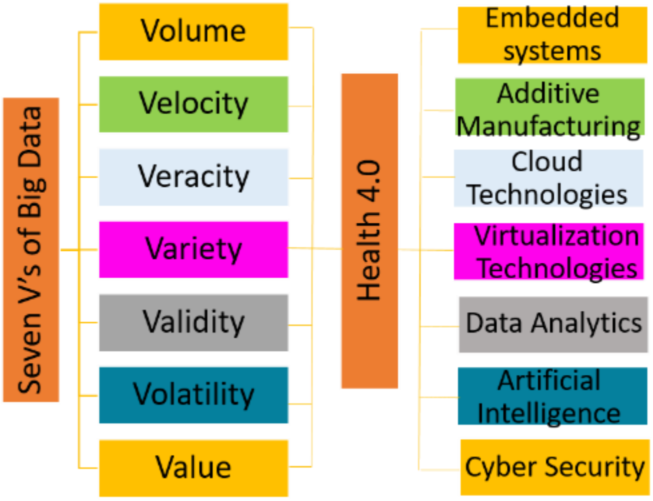


Figure5.1: Seven V's of Big Data in Association with The Latest Technology to Achieve H4.0

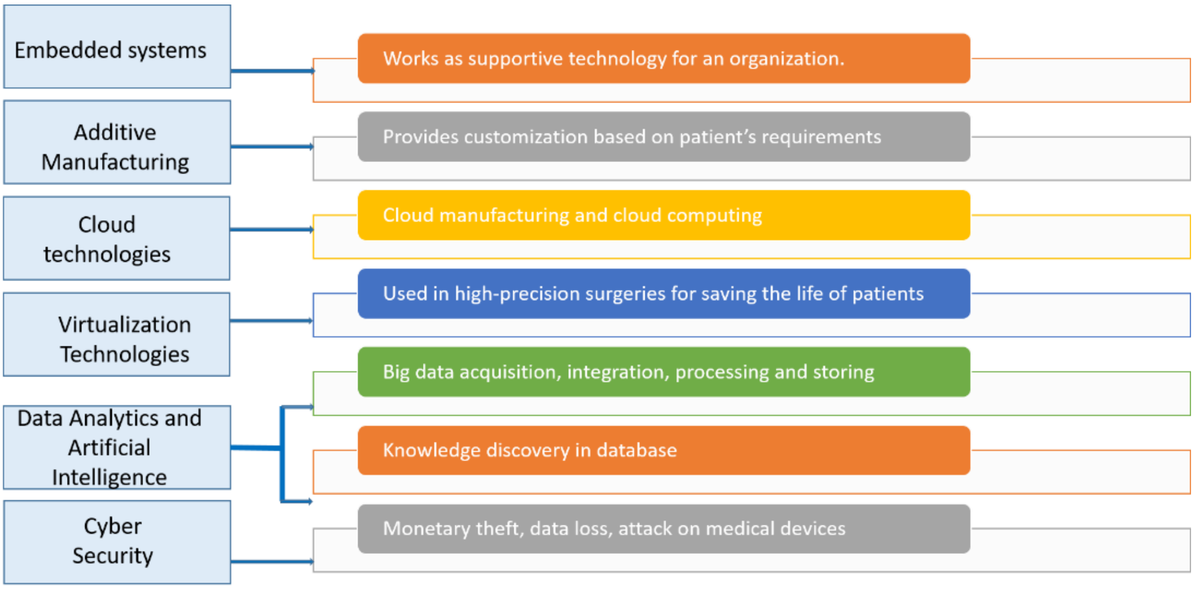

Embedded systems
Additive Manufacturing
Cloud technologies
Virtualization Technologies
Data Analytics and Artificial Intelligence
Cyber Security
Works as supportive technology for an organization.
Provides customization based on patient's requirements
Cloud manufacturing and cloud computing
Used in high-precision surgeries for saving the life of patients
Big data acquisition, integration, processing and storing
Knowledge discovery in database
Monetary theft, data loss, attack on medical devices

Figure 5.2: Use of Different Technologies to Achieve the Goal of H4.0

5.2 Proposed Work

This section covers the entire proposed work in detail.

5.2.1 Optimal Healthcare Classification Multi-Disease Diagnostic Architecture (OHCMDD)

The present work proposes a Hadoop-based OHCMDD architecture as shown in Fig. 5.3. As a test phase, the presented model has experimented on multiple disease classification. The OHCMDD architecture handles various challenges such as handling mixed-format healthcare data and combining different technologies for providing a unanimous result. The concept of health 4.0 is implemented to achieve the above-mentioned tasks and to build a proactive healthcare environment. The main objective of the architecture is the management of healthcare-related big data so that the hidden patterns can be extracted and analysed. The analysis task is also performed with the proposed D-SMO algorithm. The OHCMDD architecture consists of four layers:

1. Data Aggregation Layer
2. Data Storage Layer
3. Data Management Layer
4. Data Analysis Layer

5.2.1.1 Data Aggregation Layer

In the current scenario health-related data is generated from different sources such as mobile phones, IoT data, EHR, smart homes, etc. [154]. All the data is presented in different formats that should be properly processed which creates a big challenge in data storage and data management for extracting valuable insight. For this, data processing plays an important role to remove the unwanted data, the noise of data, and it also changes the categorical values into numerical values. All the data cleaning is performed before the next layer of the architecture handles the data efficiently and manages it to produce faster results.

5.2.1.2 Data Storage Layer

In earlier days most of the health-related data were maintained in the physical format by hospitals and it was very tough to analyse that data. Now with the advancement in storage mechanisms healthcare sectors are using the latest ways to secure the storage of data. In this study, a Hadoop-based model is utilized for data storage.

5.2.1.3 Data Management Layer

Nowadays the leading problem faced by healthcare organizations is the management of healthcare big data. The huge amount of health-related big data can be handled with the help of distributed data management techniques. Hadoop is one of the solutions as it provides a safe mechanism for storing and retrieving a big amount of data. HDFS replicates the data on various stacks of it that provide data reliability. In the OHCMDD architecture, two data placement algorithms are working on this layer to provide a safe mechanism to store and retrieve healthcare big data.

5.2.1.4 Data Analysis Layer

The processed data is now utilized for feature selection and classification tasks. The feature selection is performed on the processed dataset for selecting the most relevant features. In the present architecture, a D-SMO algorithm method is formulated where DBFS is used for feature selection and SMO is used for classification.

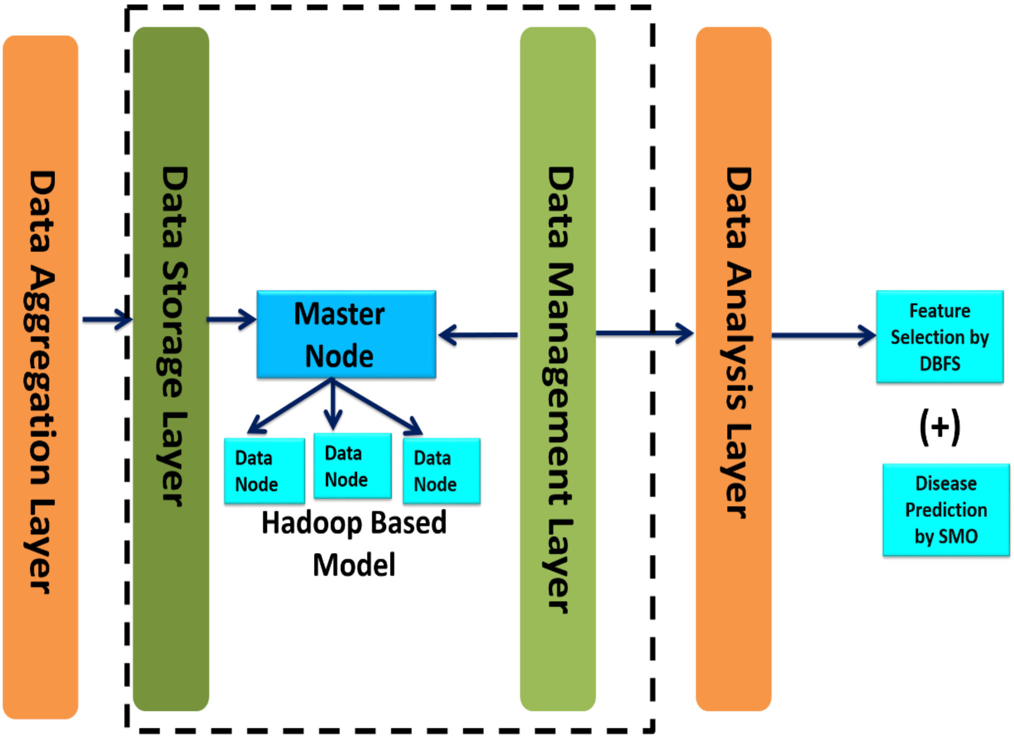


Figure 5.3: Optimal Healthcare Classification Multi-Disease Diagnostic Architecture

5.3 Proposed Approach

The correct administration of healthcare big data is the most pressing issue that the healthcare sector is confronted with. For handling this issue a distributed data management strategy is required to manage the massive amount of healthcare big data. In the present architecture two techniques are used for multiple disease detection; first is the utilization of heterogeneous Hadoop cluster using efficient data placement method, while second is the use of hybrid D-SMO algorithm for data classification.

Hadoop is based on commodity architecture, and HDFS provides a secure technique for storing and retrieving massive amounts of data by replicating it over several stacks using MapReduce. Hadoop's default architecture is based on a master-slave system [252]. In its usual context, Hadoop treats each DataNode equally while in a heterogeneous cluster, each DataNode is unique depending on its configuration. In a heterogeneous cluster, a high-speed DataNode can process its task of data processing more quickly as compared to data present in the slow node’s local disk. Based on Hadoops default strategy an efficient DataNode after finishing its job assists one of the inefficient DataNode for processing its unprocessed jobs. The performance of the Hadoop cluster is impacted by data being transferred from a sluggish DataNode to a fast

DataNode. It keeps the cluster master node highly busy by keeping track of the cluster information related to blocks and nodes. The master node is also responsible for managing all the statistical data about a node and the processes that are running on it. The master node calculates the total block number by dividing the data size (N) by the block size (M). In addition, the NameNode keeps track of the empty task slot, unprocessed jobs, and the way to assign these unprocessed jobs to a DataNode with an empty task slot. As a result, an effective data placement policy is necessary to handle the diverse behavior of DataNodes. The proposed algorithm is capable to handle the data distribution in the Hadoop cluster based on the performance of DataNode. Algorithm 5.1 calculates the Capacity Ratio (CR) for each DataNode and accordingly distributes the jobs to them. In response to it, the better performer will get more jobs in comparison to the slower one. This strategy helps the master node in maintaining the information of each DataNode and the scarce network resource is also preserved to some extent.

The working of the Hadoop model is divided into two phases. Figure 5.4 elucidates the placement strategy of a heterogeneous Hadoop cluster and Algorithm 5.1 explains the initial procedure of jobs distribution by NameNode to different DataNodes.

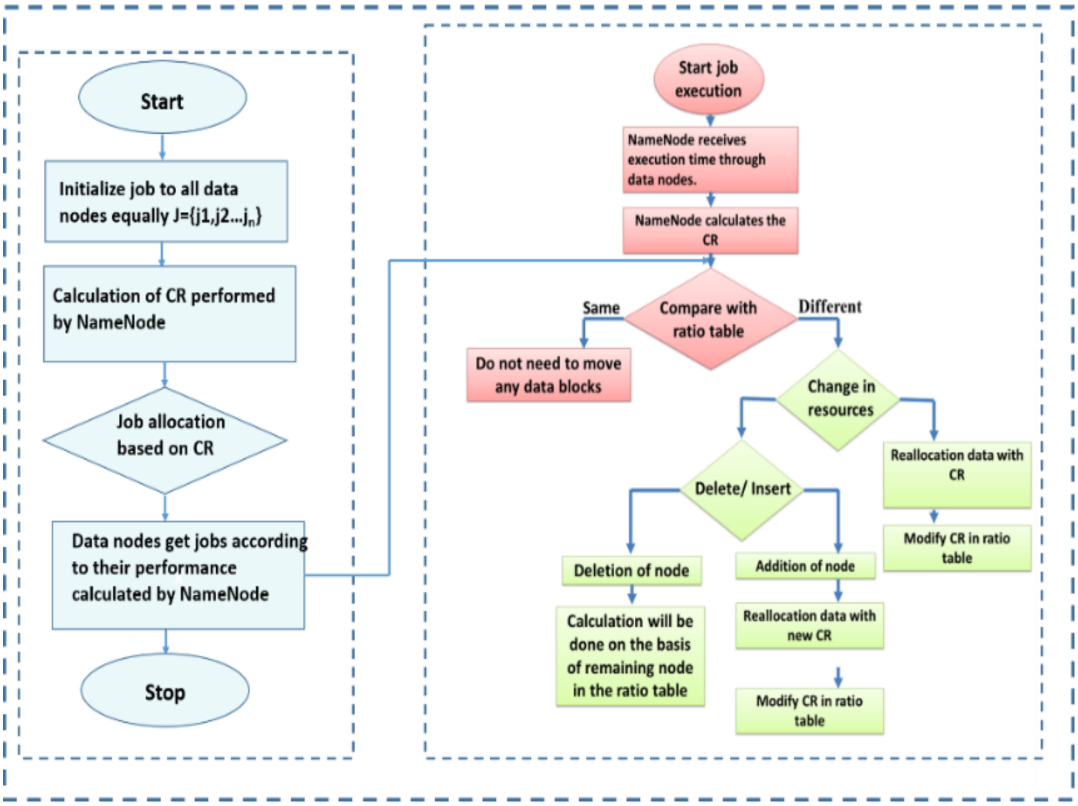

Start
Initialize job to all data nodes equally J={j1,j2...j_n}
Calculation of CR performed by NameNode
Job allocation based on CR
Data nodes get jobs according to their performance calculated by NameNode
Stop
Start job execution
NameNode receives execution time through data nodes.
NameNode calculates the CR
Same
Compare with ratio table
Different
Do not need to move any data blocks
Change in resources
Reallocation data with CR
Modify CR in ratio table
Delete/ Insert
Deletion of node
Addition of node
Calculation will be done on the basis of remaining node in the ratio table
Reallocation data with new CR
Modify CR in ratio table

Figure 5.4: Flowchart Explaining Data Placement Strategy in Hadoop Cluster by Considering Different Situations

Initially, after creating a multi-node cluster a node is set up to become the NameNode of the cluster. Now NameNode starts managing the resources of the cluster and arranging DataNodes accordingly. However, at this point, NameNode is not having any information of different DataNodes of the cluster regarding their capacity of performing different jobs. At first, when a job came for execution NameNode distributes the jobs among the DataNode equally. All DataNode are required to send their execution time to NameNode through Heartbeat messages. Based on this NameNode calculates the CR of each DataNode and records all the information in a CR table. If the same type of job arrives in the future, then NameNode will distribute the job among the DataNode according to their CR.

```
Algorithm 5.1: $Upload\ data\ on\ HDFS.$

$Input$: $JobType, = \{j_1, j_2, j_3 \dots \dots .., j_n\};$
        $DataNode, d = \{d_1, d_2, d_3 \dots \dots .., d_n\}$
$Output$: $CR_{retrieve}$

    Initial data placement on $HDFS$:
    JobType ← { $j_1, j_2, \dots . j_n$}  type of job executed by the DataNodes
    DataNodes← { $d_1$ , $d_2$, ..... $d_n$}   the DataNodes of the cluster
$CR$ ← {$\mathrm{c}_1, \mathrm{c}_2, \dots . . \mathrm{c}_\mathrm{n}$} is the ratio of the respective DataNodes obtain from the record
            $CR = \frac{work\ done}{Time\ taken}$
if   JobType is comapred with the existing records and same then same (S)
        $S = 0$
            Total block number =$\left\lceil\frac{\text{Data size}}{\text{Block size}}\right\rceil$
            Block number =$Total\ block\ number\ * \left[\frac{Node\ capacity}{\sum each\ node\ capacity}\right]$
            Node capacity← obtain from the $CR$
for   All records in the ratio table do
            When a new job arrives on $HDFS$
            And S≠0
for   The first-time data is distributed in the ratio of 1:1:1 for every new job
        Store the $CR$ of every new job in the $CR$ table
    if   $CR \neq records\ then$
            Change records based on $CR$
        Reallocate jobs to DataNodes according to present $CR$

end if
end for
end for
end if
```

The working of the second phase starts to perform horizontal scaling operation where a new DataNode is added to the cluster or an old DataNode is decommissioned from the existing cluster as shown in Algorithm 5.2. At the time of admission NameNode is not having a log for the newly admitted DataNode and the present CR table becomes futile. To overcome this problem NameNode makes the new entry in the CR table for the new DataNode. The first time after the admission of a new DataNode, NameNode provides the minimum number of jobs to the newly admitted node for execution. In the experiment, the value is fixed to 10% to the total number of jobs for the new DataNode and the remaining 90% of jobs will be distributed to the old DataNodes of the cluster in the proportion of their CR. This technique helps in improving the efficiency of the overall system. After this round, NameNode calculates the CR of the new DataNode and maintains it in the CR table. At this point, NameNode updates the values in the CR table and stores the actual ratio of new DataNode for a particular job type. Thereafter, if a new job came for execution NameNode repeats the phase one steps.

The next change appeared in the CR table at the time of DataNode's decommissioning from the existing cluster. At this point, NameNode deletes the entries of a particular DataNode from the CR table. After this NameNode distributes the data blocks present at the decommissioned node to rest of the nodes in the cluster by running a balancer. The balancer is used to redistribute the data among the DataNodes

5.3.1 D-SMO

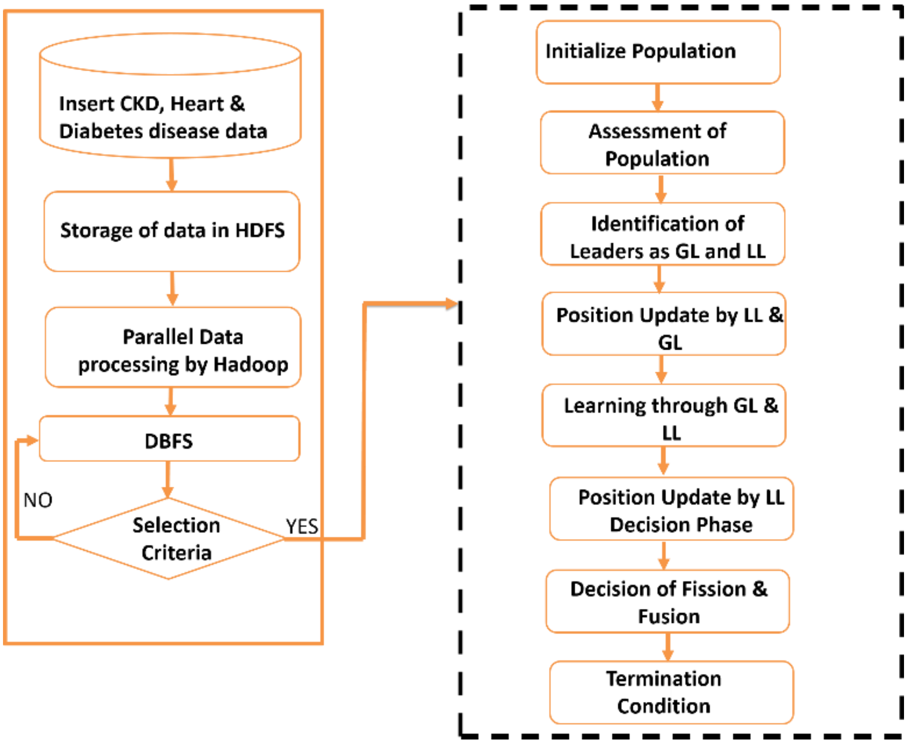
The D-SMO algorithm works on the fifth layer of the architecture. This layer performs feature selection and data classification. The data processing step is very important in upgrading the performance of the system. The goal of using a wrapper method is to select the best features by continually producing a set of features until DBFS finds the best subset. DBFS is a heuristic-driven method for evaluating the quality of a feature. Good features values for every class have minimal overlap with other classes. The DBFS method explores the contribution of every feature for assigning them appropriate rank by performing features corresponding distributions for all classes. Initially, the DBFS calculates the Probability Density Function (PDF) for all features individually in every class. After this, the ranking procedure starts for every feature depending upon the overlapping area.

```
Algorithm 5.2: $DataNode_{insert}, DataNode_{delete}$

$Input: JobType, j = \{j_n + 1\}; new\ DataNode = d_n$
$Output: CR_{retrive}$

        When a job arrives on HDFS:
        Node Number← taken from NameNode
        JobType ← {$j_1, j_2, ....j_n$} are type of job executed by the DataNodes
        Total execution time← taken from NameNode
  if    New DataNode arrives
                job $\mathrm{j_n} \rightarrow \mathrm{j_n} + 1$
            $\mathrm{X} = \sum_{\mathrm{d}=1}^{\mathrm{n}} \mathrm{CR} + \mathrm{d_{n+1}} = 1$
      Amount of data assigned to new DataNode for first time is 0.1 or 10%
  else
          $X$= Total number of jobs
          $X = CR_1d_1 + CR_2d_2 + CR_3d_3 + CR_nd_n$
          $\sum_{d=1}^{n+1} CR = 1$
    $CR$ ← set 10% for new node and reaming 90% for the existing node in their obtained $CR$
        Node capacity← obtained from $CR$ of the new node
          Add job type CR of new node in $CR$ table
    if    DataNode is decommissioned from the cluster
          //Entries of particular DataNode deleted from the cluster and then run the Balancer
      Run← $\%start - balancer.sh$
      NameNode← delete the entries of the particular DataNode from the cluster
      Reallocation of the data blocks among the existing DataNodes
  Close if
  Close if
```

However, there are two methods present to calculate the PDF, the first is the parametric approach and the second is the non-parametric approach. In the present study parametric approach is utilized.

$$p(x) \cong \frac{k}{NV} \tag{5.1}$$

In Equation 5.1 $p(x)$ is the calculated PDF for instance (x), and V is the surrounded volume around x. The total number of instances are represented by N and k are the instances inside V. To estimate the precise PDF, the value of N should be increased and the value of V decreased. The combination of DBFS and SMO in the D-SMO algorithm helps healthcare professionals and patients to foresee and analyse their medical data. The D-SMO model will achieve a good classification rate with lesser features set that lead to attaining optimal performance. Moreover, the flowchart of Hadoop based D-SMO algorithm is provided in Fig. 5.5.



Insert CKD, Heart & Diabetes disease data
Storage of data in HDFS
Parallel Data processing by Hadoop
DBFS
NO
Selection Criteria
YES
Initialize Population
Assessment of Population
Identification of Leaders as GL and LL
Position Update by LL & GL
Learning through GL & LL
Position Update by LL Decision Phase
Decision of Fission & Fusion
Termination Condition

Figure 5. 5: Flowchart of Hadoop Based D-SMO Algorithm

5.3.2 Spider Monkey Optimization

SMO algorithm comes under the meta-heuristic algorithms and it is inspired by the foraging behavior of SM. The foraging nature of SM depends upon the fission-fusion social structure. The characteristics of the SMO algorithm are based on the social organization of a group. In this group of SM all the decisions of splitting or combining the group are taken by the female leader. About the algorithm, the existing situation of food scarcity represents no improvement in the solution. As the SMO algorithm is a swarm-based algorithm and every small group of SM should have a minimum number of members in it. So, in the future for fission all groups should have a minimum number of members, if at least one group is not having a minimum number of members, then fusion will take place. In the SMO algorithm, the potential solution is represented as a Spider Monkey.

5.3.2.1 Foraging Behavior

The foraging behavior of SM is represented in Fig. 5.6. The SM generally prefers to live in a single group known as the parent group. The decision of splitting and uniting the group is based on the availability of the food. Spider Monkey communicates by using different gestures, whooping, and positions. The composition of the group is the dynamic property of SM's group structure. Commonly, 40-50 SM live in a group and forage in small groups based on the availability of food in different directions. They tend to share their foraging experience in the night and all the forage routes are decided by female SM. The individuals of the SM group may not be noticed as closely at one location due to their mutual tolerance. The classification task of CKD, heart, and Diabetes by applying the SMO algorithm involves the following processes:

- Local Leader phase (LLP)
- Global Leader phase (GLP)
- Local Leader Learning phase (LLLP)
- Global Leader Learning phase (GLLP)
- Local Leader Decision phase (LLDP)
- Global Leader Decision phase (GLDP)

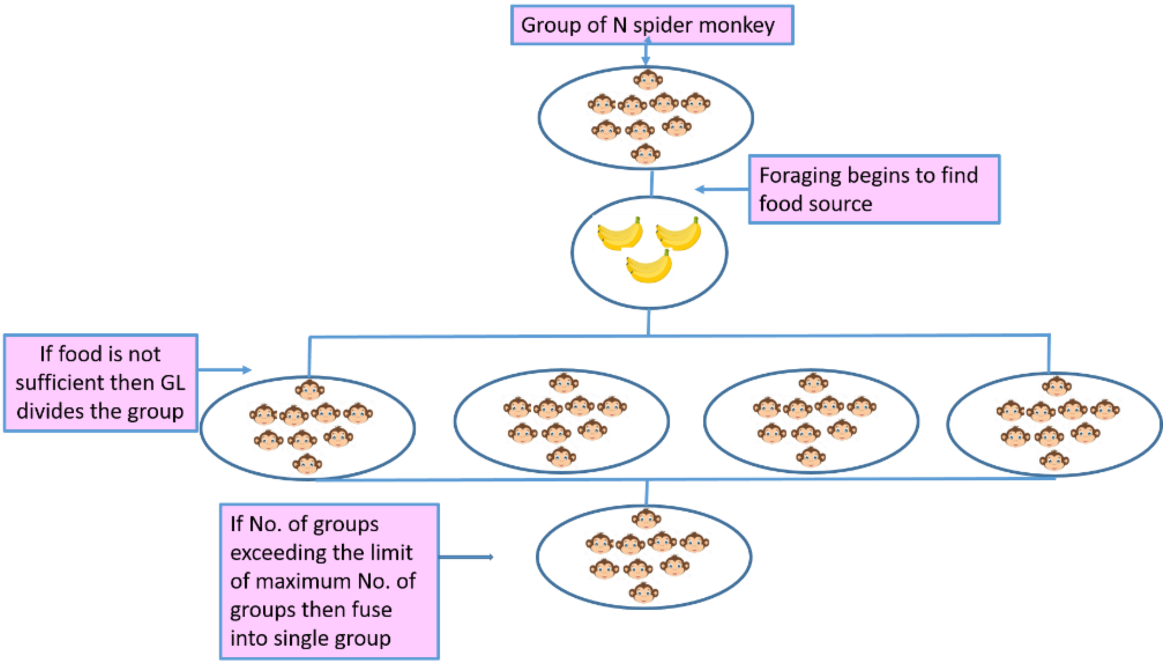

Group of N spider monkey
Foraging begins to find food source
If food is not sufficient then GL divides the group
If No. of groups exceeding the limit of maximum No. of groups then fuse into single group

Figure 5.6: Flowchart showing the Foraging Behavior of SM

Initialization: In the beginning, SMO performs a uniform distribution of the initial swarm of N SM. Equation 5.2 represents the initialization phase of SMO where SM_n is the n^{th} SM in the swarm. The initialization of SM_n is performed as:

$$SM_{nq} = SM_{minq} + UR(0,1) * (SM_{maxq} - SM_{minq}) \quad (5.2)$$

Here, the lower bound is SM_{minq} and the upper bound is SM_{maxq} in the q^{th} dimension of the search space. The range of uniformly distributed random numbers used is (0,1) represented by UR $(0,1)$.

- **Local Leader Phase (LLP):** In the SMO algorithm the LLP is an imperative phase. In this phase, all SM get a chance to update their position. The modification in the position of SM depends on the experience of local group members and LL. The fitness value of SM is calculated for its new position and if the fitness value is higher than the old fitness value is updated otherwise it remains the same. The equation of position update is Equation 5.3.

$$SMnew_{nq} = SM_{nq} + UR(0,1) * \left(LL_{mq} - SM_{nq}\right) + UR(-1,1) * (SM_{rq} - SM_{nq}) \quad (5.3)$$

Where, SM_{nq} is the q^{th} dimension of the n^{th} $SM.$ LL_{mq} represents the location of the local group leader of m^{th} group. SM_{rq} denotes the q^{th} dimension of the randomly selected n^{th} SM from the m^{th} group. UR denotes the uniformly distributed random number ranges between (-1, 1) and $r \neq n$. The whole position update procedure of LLP is explained in Algorithm 5.3.

- Global Leader Phase (GLP): The GLP starts after the completion of LLP. In GLP to update the position of SM different steps are required like neighbouring experience, global leader knowledge, and its own persistence as explained in Algorithm 5.4. The position update Equation 5.4 for GLP is as follows

- $SMnew_{nq} = SM_{nq} + UR(0,1) * \left(GL_q - SM_{nq}\right) + UR(-1,1) * (SM_{rq} - SM_{nq})$ (5.4)

Algorithm 5.3: *Position update method for LLP*

Input: *Initial swarm of N SM*

Output: *Fitness value of LL SM*

```
for  Every member $SM_n \in M^{th}$ group do
    for  each $q \in \{1, \dots \dots, D\}$ do
        if  $UR\ (0,1) \geq pr$ then
            $SMnew_{nq} = SM_{nq} + UR(0,1) * \left(LL_{mq} - SM_{nq}\right) + UR(-1,1) * \left(SM_{rq} - SM_{nq}\right)$
        else
            $SMnew_{nq} = SM_{nq}$
        end if
    end for
end for
```

Algorithm 5.4: *Position update method for GLP*

Input: *Initial swarm of N SM*

Output: *Fitness value of GL SM*

```
$count = 0$
while  Count < maximum group size do
    for  Each member $SM_n \in$ group do
        if  $UR\ (0,1) < prob_n$ then
            $count = count + 1$
            random selection $q \in \{1, \dots \dots, D\}$
            random selection $SM_r \in$ group $r \neq n$
            $SMnew_{nq} = SM_{nq} + UR(0,1) * \left(GL_q - SM_{nq}\right) + UR(-1,1) * \left(SM_{rq} - SM_{nq}\right)$
        end if
    end for
end while
```

Where GL_q is the position of the GL in the q^{th} dimension. The equation is divided into three parts: the first part is about the persistence of the current SM, the second part represents the attraction of current SM towards GL, and the third part is about the stochastic behavior of the algorithm. The second part of Equation 5.4 is used to improve the experience of already recognized search space, while the last phase is used to avoid the chance of getting stuck in local optima. In this phase, the SM get a chance to update its position based on the value of $Prob_n$. Where $Prob_n$ is the selection probability of the n^{th} SM to be selected in the GL phase. Therefore, the SM who's having a better location can get more chances or increased probability to make themselves better in comparison to the less fit SM. The probability of the GL is calculated by using any one solution out of the two given in Equation 5.5 and 5.6.

$$Prob_n = \frac{fitness_i}{\sum_{i=1}^{N} fitness_i} \text{ or} \tag{5.5}$$

$$Prob_n = 0.9 * \frac{fit_i}{\max\_fit} + 0.1 \tag{5.6}$$

- **Global Leader Learning Phase (GLLP):** In this phase, a greedy selection is performed to update the location of GL from the entire swarm. The optimum location is identified for an SM who will become the GL of the swarm. Additionally, a counter is associated with the GL named as Global Limit Count (GLC) to check the change in the position of a GL. The GLC value is incremented to 1 if no updates are encountered else value remains 0. The GLC is associated with GL to be compared with Global Leader Limit (GLL).
- **Local Leader Learning Phase (LLLP):** The method of greedy selection is also applied to the segment of LLLP of the SMO algorithm to update the position of LL among the group members. Every group finds its LL by updating the value of Local Limit Count (LLC). Based on the optimal location of the LL the value of the LLC is updated. If there is no change in the position of the LL, then the counter associated with LL is incremented by 1 else it is set to 0.

- **Local Leader Decision Phase (LLDP):** All the selection work of GL and LL has been performed before LLDP. The working of LLDP starts when a LL does not update its LLL to a particular verge. In respect to this, all the members of a particular group modify their position randomly or by using the experience of GL as explained in Equation 5.7 by applying pr known as perturbation rate. Algorithm 5.5 explains the entire working of LLDP.

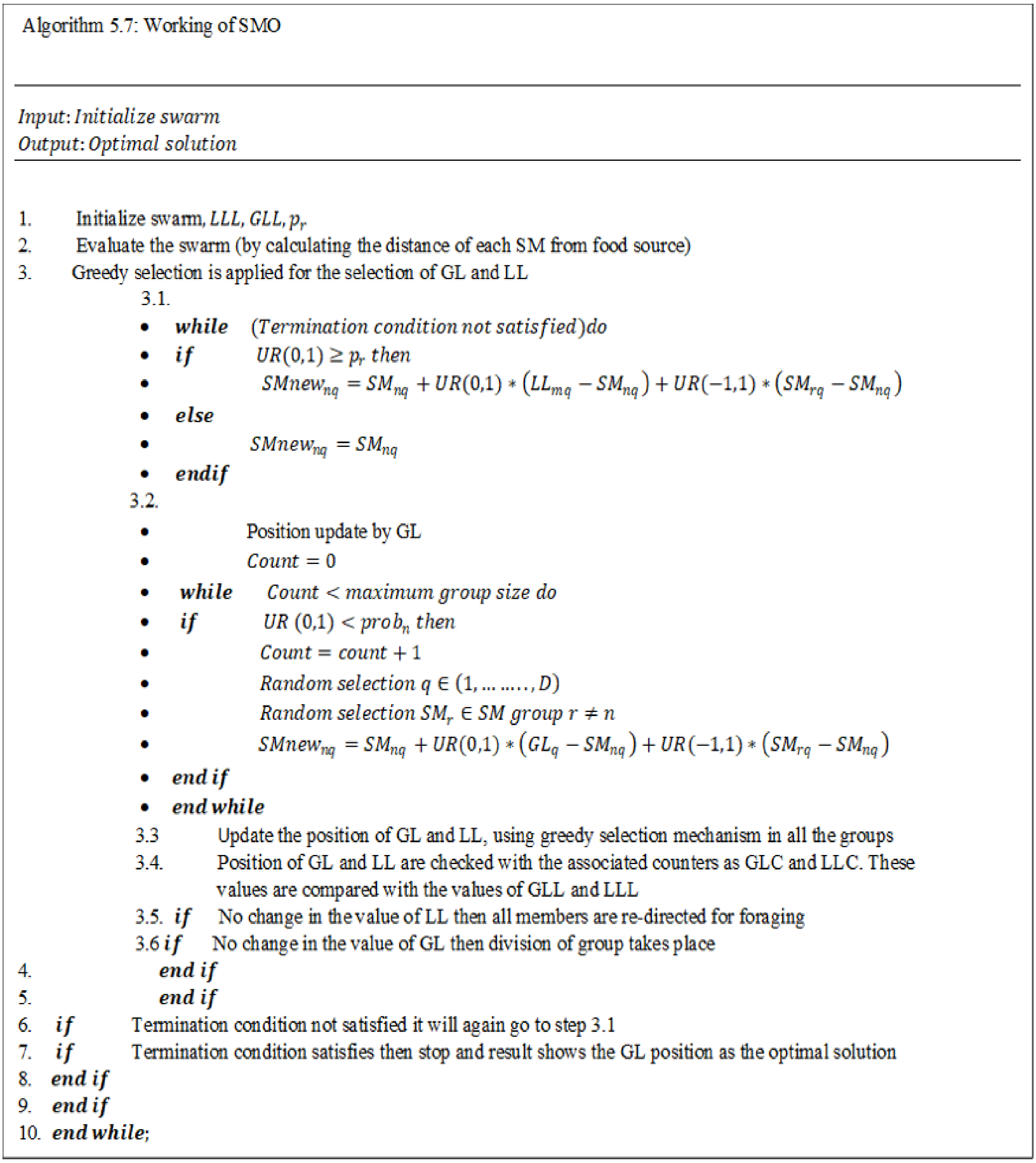
$$SMnew_{nq} = SM_{nq} + UR(0,1) * \left(GL_q - SM_{nq}\right) + UR(0,1) * (SM_{rq} - LL_{mq}) \tag{5.7}$$

Algorithm 5.5: Working of $LLDP$

```
$Input: LLC > LLL$
$Output: LLD$

$\textbf{if}$  $LLC > LLL\ then$
    $LLC = 0$
    $\textbf{for}$  $each\ q\ \in (1, \dots, D\}do$
        $\textbf{if}$  $UR(0,1) > pr\ then$
            $SMnew_{nq} = SM_{minq} + UR(0,1) * \left(SM_{maxq} - SM_{minq}\right)$
        $\textbf{else}$
            $SMnew_{nq} = SM_{nq} + UR(0,1) * \left(GL_q - SM_{nq}\right) + UR(0,1) * \left(SM_{rq} - LL_{mq}\right)$
        $\textbf{end if}$
    $\textbf{end for}$
$\textbf{end if}$
```

Algorithm 5.6: The working of $GLDP$

```
$Input: GLC > GLL$
$Output: Fission - fusion$

$\textbf{if}$  $GLC > GLL\ then$
    $GLC = 0$
    $\textbf{if}$  $Total\ number\ of\ groups < maximum\ number\ of\ groups\ then$
        $Division\ of\ swarm\ takes\ place$
    $\textbf{else}$
        $\mathrm{R}ecombine\ all\ groups\ to\ form\ parent\ group$
    $\textbf{end if}$
    $\mathrm{U}pdate\ the\ position\ of\ LL$
$\textbf{end if}$
```



Algorithm 5.7: Working of SMO

Input: Initialize swarm
Output: Optimal solution

```
1.  Initialize swarm, $LLL, GLL, p_r$
2.  Evaluate the swarm (by calculating the distance of each SM from food source)
3.  Greedy selection is applied for the selection of GL and LL
        3.1.
        • while  $(Termination\ condition\ not\ satisfied)do$
        • if      $UR(0,1) \geq p_r$ then
        •             $SMnew_{nq} = SM_{nq} + UR(0,1) * (LL_{mq} - SM_{nq}) + UR(-1,1) * (SM_{rq} - SM_{nq})$
        • else
        •             $SMnew_{nq} = SM_{nq}$
        • endif
        3.2.
        •             Position update by GL
        •             $Count = 0$
        • while   $Count < maximum\ group\ size$ do
        • if      $UR\ (0,1) < prob_n$ then
        •             $Count = count + 1$
        •             Random selection $q \in (1, \dots \dots, D)$
        •             Random selection $SM_r \in SM\ group\ r \neq n$
        •             $SMnew_{nq} = SM_{nq} + UR(0,1) * (GL_q - SM_{nq}) + UR(-1,1) * (SM_{rq} - SM_{nq})$
        • end if
        • end while
        3.3      Update the position of GL and LL, using greedy selection mechanism in all the groups
        3.4.     Position of GL and LL are checked with the associated counters as GLC and LLC. These
                 values are compared with the values of GLL and LLL
        3.5. if  No change in the value of LL then all members are re-directed for foraging
        3.6 if   No change in the value of GL then division of group takes place
4.          end if
5.          end if
6.  if      Termination condition not satisfied it will again go to step 3.1
7.  if      Termination condition satisfies then stop and result shows the GL position as the optimal solution
8.  end if
9.  end if
10. end while;
```

- **Global Leader Decision Phase (GLDP):** The GL decision is based on the value of GLL, if the GL does not update its location up to the predefined verge of GLL then the GL divides the population into smaller groups. To check the group splitting process, the GLC is compared with the GLL if it greater than the GLL then the value of GLC is set to zero. The number of the existing group is compared with the maximum number of the pre-defined group if it is less than the maximum number the GL further divides the group otherwise forms a single parent group as shown in Algorithm 5.6. Algorithm 5.7 explains the full working of the SMO process and the parameters that control the process of SMO are:
- Perturbation rate(pr)
- Maximum number of groups
- Value of GLL
- Value of LLL

5.4 Experimental Setup

This section explains the experimental setup of OHCMDD architecture to predict multiple diseases. The OHCMDD architecture works on five layers and every layer is dedicated to perform a particular task. The experimental environment is real where a Hadoop cluster of five nodes is used. In the cluster, one DataNode plays the role of NameNode, which acts as a master node, and the remaining four nodes become the slave nodes of the cluster. All nodes in the cluster use the same version of Hadoop i.e. 1.1.2 and Ubuntu 14.04 as the operating system. Although the disk space and internal memory vary for all nodes of the cluster. The configuration of every DataNode is shown in Table 5.1.

Hadoop: Hadoop is open-source software that provides the functionality to efficiently store and process big datasets. Hadoop is a combination of four modules: MapReduce, YARN, HDFS, Hadoop common. Hadoop performs distributed processing among the cluster nodes.

- **MapReduce:** To store structured data on HDFS, MapReduce is utilized for parallel processing. It is used in the Hadoop cluster to provide scalability with thousands of DataNodes. MapReduce is used in Hadoop for distributing large amounts of data over several clusters. There are two critical steps in MapReduce task processing: Map and Reduce. Each step in the file system contains input and output key values. The

MapReduce task organizes unstructured data according to the user's specifications. After that, the structured data will be appropriately preserved on HDFS for later processing.

- **R:** R is also an open-source and powerful language used for data analytics. R supports a big range of mathematical and statistical libraries. The combination of R with Hadoop scales their performance by transforming simple data analytics to strong or big data analytics with better visualization features. To use the functionality of R on the Hadoop cluster the RHadoop is installed along with its packages on every node separately. Generally, RHadoop comes in a collection of three packages namely: rhdfs, rmr and rhbase. The rmr package provides MapReduce functionality which directly calls the MapReduce API for performing MapReduce jobs on the Hadoop cluster. The rhdfs is used to do the task of HDFS jobs on Hadoop by calling HDFS API. In the present architecture the version of Hadoop used is 1.1.2, R is 3.1.0, rmr2 is 3.1.0 and rhdfs is 1.0.8 used.

Table 5.1: Configuration of the DataNodes Employed in the Cluster

Node	CPU	Memory	Disk	Operating system	Hadoop version	R	Java
Master node	Intel Core2duo	4 GB	43.7 GB	Ubuntu 14.04	1.1.2	3.1.0	JDK 7
Slave node 1	Intel Core2duo	4 GB	56.4 GB	Ubuntu 14.04	1.1.2	3.1.0	JDK 7
Slave node 2	Intel Core2duo	4 GB	99.0 GB	Ubuntu 14.04	1.1.2	3.1.0	JDK 7
Slave node 3	Intel Core2duo	4 GB	45.3 GB	Ubuntu 14.04	1.1.2	3.1.0	JDK 7
Slave node 4	Intel Core I5	8.0 GB	499 GB	Ubuntu 14.04	1.1.2	3.1.0	JDK 7

5.5 Results and Discussion

5.5.1 Comparison of the Proposed Model with Existing Features and Proposed Features

The model has experimented with CKD, heart and diabetes disease datasets from the UCI repository. The investigation was carried out after selecting the best features from the dataset of different diseases using DBFS. To handle the different disease data efficiently the heterogeneous Hadoop cluster is used with the R tool. The importance of feature selection is examined in Table 5.2 where existing features are compared with the proposed features. Here, the performance metrics like sensitivity, Precision, accuracy and F1-score were evaluated and shows positive performance.

Figure 5.7 (A) shows the selected features for the CKD dataset, Fig. 5.7 (B) presents the vital features for heart disease and Fig. 5.7 (C) shows the significant features of Diabetes disease. The DBFS method selects a different number of features for all three databases of various diseases in 10 iterations by 10-fold cross-validation. An average of 12 features was selected for CKD, 13 for heart disease and 5 for Diabetes.

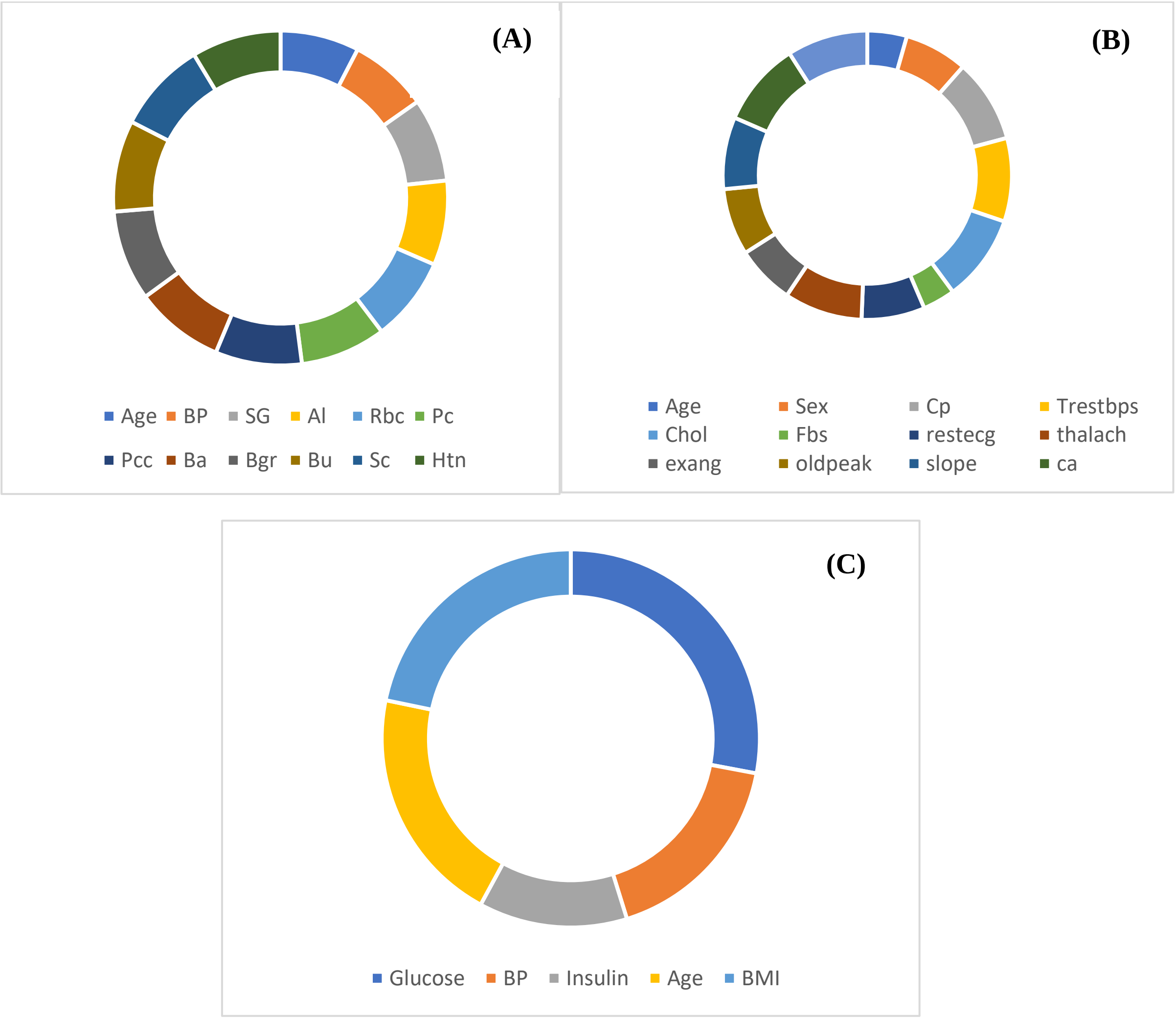


Figure 5.7: (A) shows the Important Features of CKD (B) Presents the Important Features of Heart Disease (C) Gives the Vital Features of Diabetes Disease As Selected By DBFS Algorithm

Table 5.2: Presents the Result for CKD, Diabetes and Heart Disease with the Existing Features of the Dataset and the Proposed Features

S.No	**Database**	**With existing features**				**Proposed features**			
	Diseases	Accuracy	Precision	Sensitivity	F1	Accuracy	Sensitivity	Precision	F1
1.	CKD	94.1	87.5	90.4	88.9	98.30	100	96.80	98.3
2.	Diabetes	91.2	92.8	91.7	92.2	94.90	92.3	94.20	93.2
3.	Heart	84.5	84.4	89.3	86.8	99.70	90.2	92.30	91.2

5.5.2 Comparison of Classification Algorithm

The reduced features are identified as the important set of features to produce a good result from the classification phase. This set of reduced features are implemented with D-SMO algorithm and compared with existing methods such as ACO, SMO, DT and RF. The classification results of various algorithms are shown in Table 5.3 and Fig. 5.8 (A), (B), (C) and (D) for heart disease, CKD and Diabetes.

Table 5.3: Analyzing the Accuracy, Precision, Sensitivity and F1-score of CKD, Diabetes and Heart Disease with Different ML Algorithms

Evaluation matrices	**Disease**	**D-SMO (%)**	**ACO (%)**	**SMO (%)**	**DT (%)**	**RF (%)**
Accuracy	CKD	98.30	92.70	94.1	90.50	90.90
	Diabetes	94.90	89.50	91.2	84.34	86.28
	Heart Disease	93.70	86.80	84.5	83.49	88.14
Precision	CKD	96.80	88.15	87.50	85.72	86.31
	Diabetes	94.20	79.10	92.83	70.10	82.83
	Heart Disease	92.30	83.12	84.49	79.46	78.83
Sensitivity	CKD	1	88.50	90.40	89.20	85.40
	Diabetes	92.3	87.90	91.70	85.60	87.7
	Heart Disease	90.2	87.20	89.30	87.60	89.9
F1-Socre	CKD	98.3	88.31	88.92	87.42	85.85
	Diabetes	93.2	83.26	92.26	77.07	85.19
	Heart Disease	91.2	85.11	86.82	83.33	84

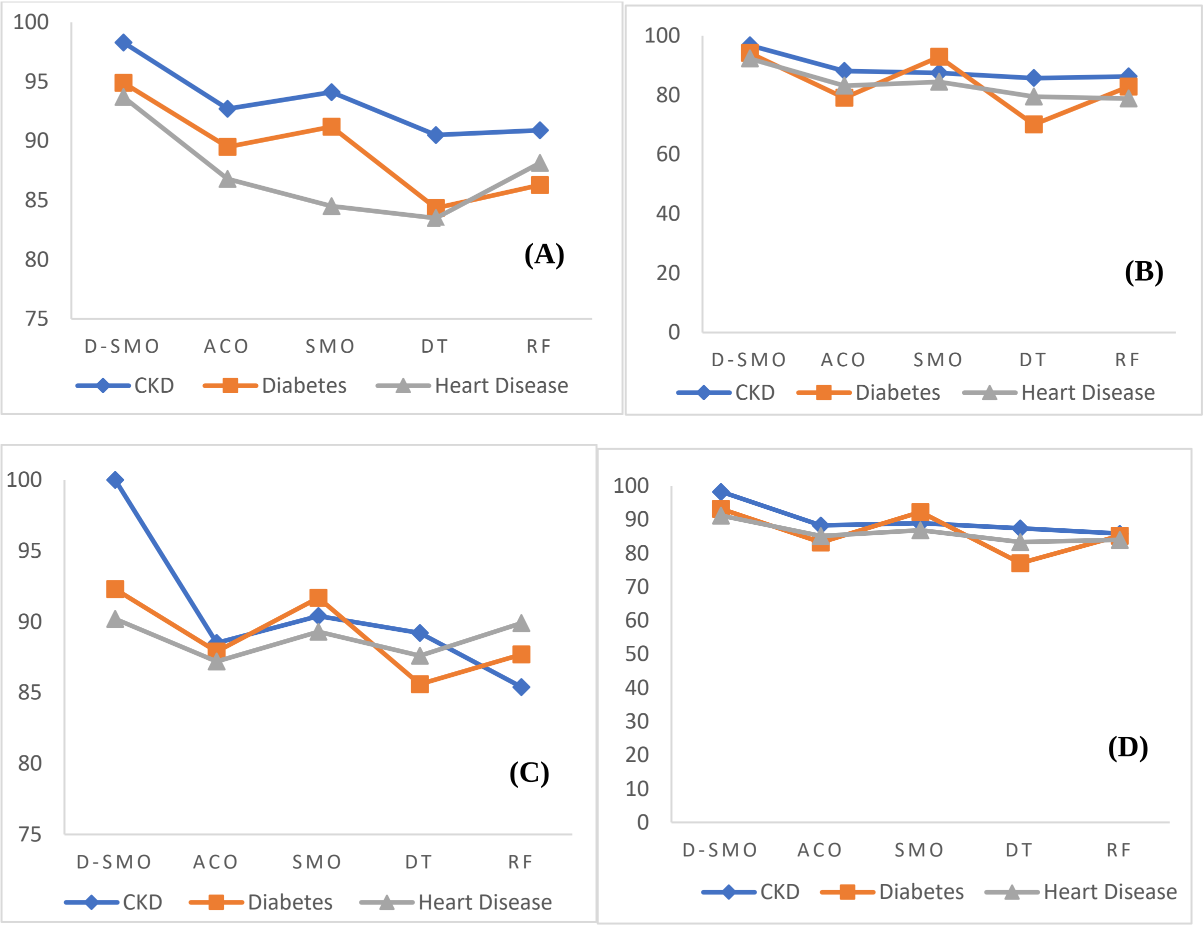


Figure 5.8: Comparison of D-SMO, ACO, SMO, DT and RF with respect to (A) Accuracy, (B) Precision, (C) Sensitivity, and (D) F1-Score.

5.5.3 Comparative Analysis with Existing Studies

This section presents the comparative study between the state-of-the-art methods and different aspects of the proposed architecture. The OHCMDD architecture is divided into two parts the first part is dedicated to data storage and management and the second part performs the classification work. The data management is handled by the Hadoop-based model where efficient data placement algorithms are employed to manage different disease datasets. The performance of the proposed hybrid classification algorithm along with the data distribution scheme is evaluated and compared with the existing work [108, 110, 111, 244].

In the study presented by Youvraj et al. [108], the author uses a Hadoop-based model for data management and ML algorithms are used for the classification of Diabetes disease. However, the proposed architecture uses Hadoop for data distribution and management by improving the flaws in the default architecture of Hadoop. Along with this, the implementation of the hybrid algorithm is applied for disease prediction. The improvement in the results can be seen in Fig. 5.9.

In the study presented by Abdelaziz et. al. [110], an ML-based model is utilized to provide better healthcare services in a cloud environment. The architecture is implemented on the cloud to optimize the selection process of virtual machines using PSO. The virtual machine selection is performed to maximize the utilization of the cloud to process the medical requests coming from different stakeholders. In addition, the author uses a hybrid model for Kidney disease diagnosis where LR is used for finding critical factors and NN is used for prediction. As shown in Fig. 5.9 the OHCMDD architecture proves to be more effective as it works in the real environment of the Hadoop cluster for data management and prognosis of multiple diseases done with D-SMO.

The study proposed by Elhoseny et al. [111], has proposed a hybrid algorithm for CKD. The author used the combination of DFS for extracting the features of CKD while ACO is used for prediction. As shown in Table 5.4 the performance of the D-SMO algorithm is better than the algorithm presented in [111]. Along with this, the OHCMDD architecture is also doing data management in the heterogeneous environment.

Moreover, in the study presented by Xiaohua et al. [244], uses a different combination of algorithms on the Diabetes dataset. The combination of genetic algorithm, PSO, and Harmony search is used with K-means for feature extraction and KNN used for prediction of Diabetes. Their result shows that the best performer was Harmony-K-means with KNN. The comparison based on accuracy, precision, recall, and F1-score indicates that the performance of the D-SMO algorithm is better than the combination presented in [244]. Therefore, the comparative analysis of the performance evaluation shows that the OHCMDD architecture and D-SMO algorithm are more effective and able to provide an improved and better solution for the healthcare system.

The performance analysis of the present D-SMO algorithm on comparing with other studies [108, 110, 111, 244] as given in Table 5.4 is improved in terms of accuracy ranging from 0.82% to 9.587%, 1.82% to 4.17%, and 3.42% to 4.34% as a minimum to maximum range for CKD, diabetes and heart disease, respectively. The precision value is upgraded from 1.25% to 10.75%, 0.10% to 7.65%, 1.09% to 2.66% for CKD, diabetes, heart disease respectively. While the improved value for sensitivity is 0.61% to 14.70%, 1.30% to 4.13% and 0.55% to 2.58% respectively for CKD, diabetes, heart disease. The betterment in F-1 score is also noticed as 6.61% to 14.70%, 1.30% to 4.13% and 0.55% to 2.58% for CKD, diabetes, heart disease respectively.

Table 5.4: Analysing the Performance of OHCMDD Architecture's Prediction for CKD, Diabetes and Heart Disease with Existing Studies

Disease	Accuracy	Precision	Sensitivity	F-1
CKD	98.3	96.8	100	98.3
[Proposed] Diabetes	94.9	94.2	92.3	93.2
Heart	93.7	92.3	90.2	91.2
[108] CKD	89.7	87.4	84.2	85.7
Diabetes	92.9	92	89	90
Heart	90.2	91.3	90.1	90.7
[110] CKD	97.5	95.6	100	97.7
Diabetes	93.2	94.1	90	92
Heart	90.6	89.9	90.1	89.9
CKD	95.7	94.2	96.8	95.4
[111] Diabetes	91.1	90.7	91.2	90.9
Heart	89.8	90.2	87.7	88.9
[244] CKD	94.7	91.3	90.8	91.0
Diabetes	92.2	87.5	91.7	89.5
Heart	90.4	91	89.2	90.1

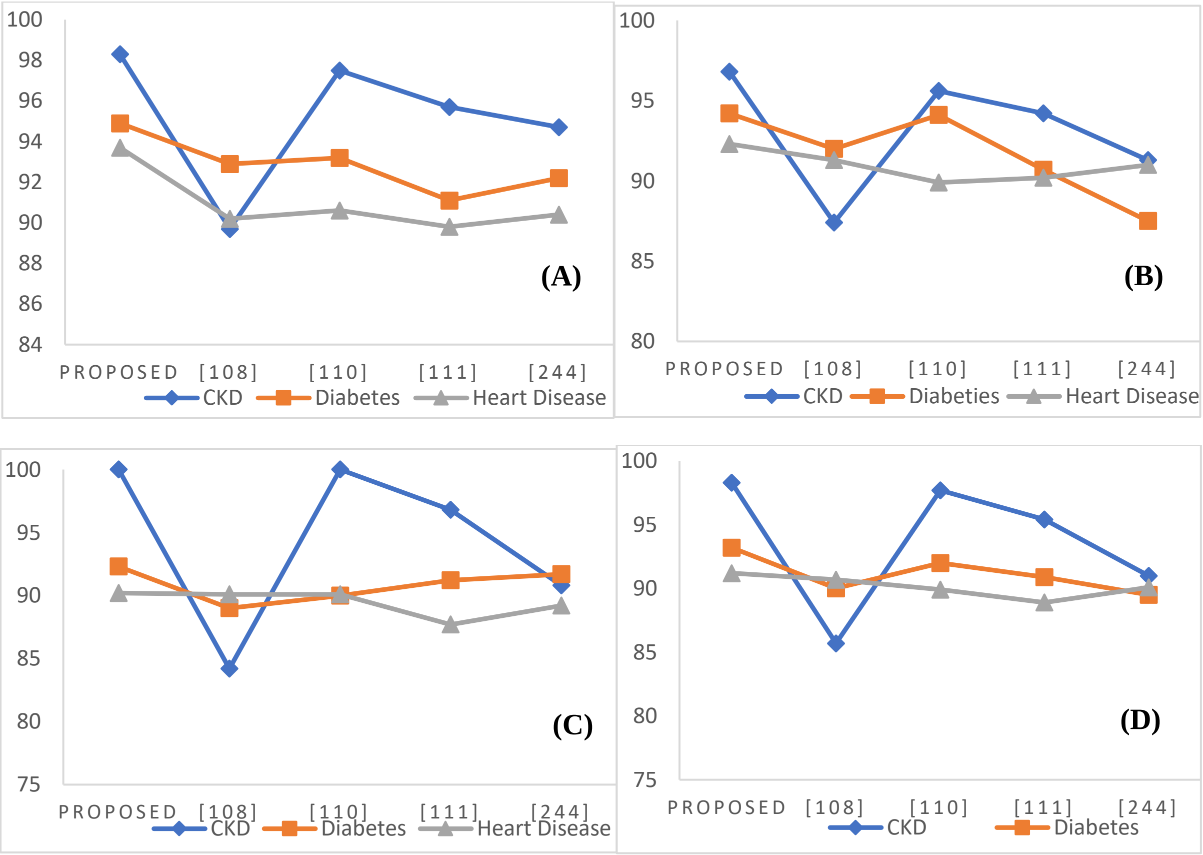


Figure 5.9: Performance Comparison of OHCMDD Architecture with [108, 110, 111, 244] (A) Accuracy, (B)Precision, (C) Sensitivity, and (D)F-1

5.6 Summary

A new Hadoop-based optimal healthcare classification multi-disease diagnostic architecture has been successfully devised to improve healthcare services. This OHCMDD architecture worked on five layers. The Hadoop cluster environment is created with R to handle the dataset of multiple diseases. It also handles the issues of horizontal scaling in the Hadoop cluster. Along with this the present study also presents the intelligent classification and prediction system for healthcare namely the D-SMO algorithm. The D-SMO algorithm is a combination of DBFS and SMO algorithms and it jointly performs feature selection, removing irrelevant features and SMO-based prediction. Three benchmark disease datasets are used to evaluate the efficiency of the D-SMO algorithm. Moreover, a comparison study was also performed with the existing methods. The comparison results show that the OHCMDD architecture outperformed the other existing methods with improved data distribution techniques and classification results generated by D-SMO.

In addition to the proposed architecture that handles multiple diseases, various other aspects are also associated in the healthcare industry that comes under the umbrella of H4.0 and may be further improvised in order to provide a better healthcare environment. In the upcoming chapter the H4.0 architecture was developed along with the latest technologies required for it and the road map is also presented to achieve health 4.0.

CHAPTER 6

eHealth Architecture for Health 4.0

This chapter presents the eHealth architecture for the implementation of Health 4.0 (H.40). Along with this, a road map is also provided to achieve the target of H4.0 by implying the principles of Industry 4.0 (I4.0).

6.1 Introduction

The H4.0 approach mortgaged the ‘‘4.0’’ attribute from I4.0. The health sector is the adequate technological conjunction of the features of the Fourth Industrial Revolution. The four generations are mechanization, automation, computerization, and informatization as depicted in Fig. 6.1. H4.0 incorporates the principles of I4.0 for the digitization of laboratories and to implement automation in numerous processes used in the general health sector and hospitals. The health sector is facing lots of challenges in terms of services such as the management of different disease datasets that develop over the years. H4.0 is used to improve the efficiency of physicians by enhancing their speed for exploring patients’ data and, enabling them to optimize the resources so that patients’ health can be improved. This all can be done by extending the boundaries of innovation with the collaboration of the Internet of Services (IoS) and IOT [253].

The genesis of H4.0 spreads its dawn in the third millennium with the emergence of I4.0. It is the only industrial revolution rooted in a new technological phenomenon known as digitalization. This digitalization enables us to develop a new virtual world to be used as steering to steer the physical world. Today’s industries aim to connect all production means so that their interaction can be done in real-time.

To achieve this target, Reference Architecture Model Industry 4.0 (RAMI 4.0) is used with I4.0 and the purpose of RAMI 4.0 is to make sure that all the involved participants in I4.0 have a common framework to understand each other. The three-dimensional map of RAMI 4.0

shows the structured method for the deployment of I4.0. The RAMI 4.0 architecture helps the factories to present themselves as an interconnected network of smart products with the connected world. The main principle of I4.0 is to maximize the efficiency of production by minimizing its cost and it can be achieved efficiently with the help of RAMI 4.0. This technological concatenation also helps to achieve H4.0 where all the involved participants understand each other.

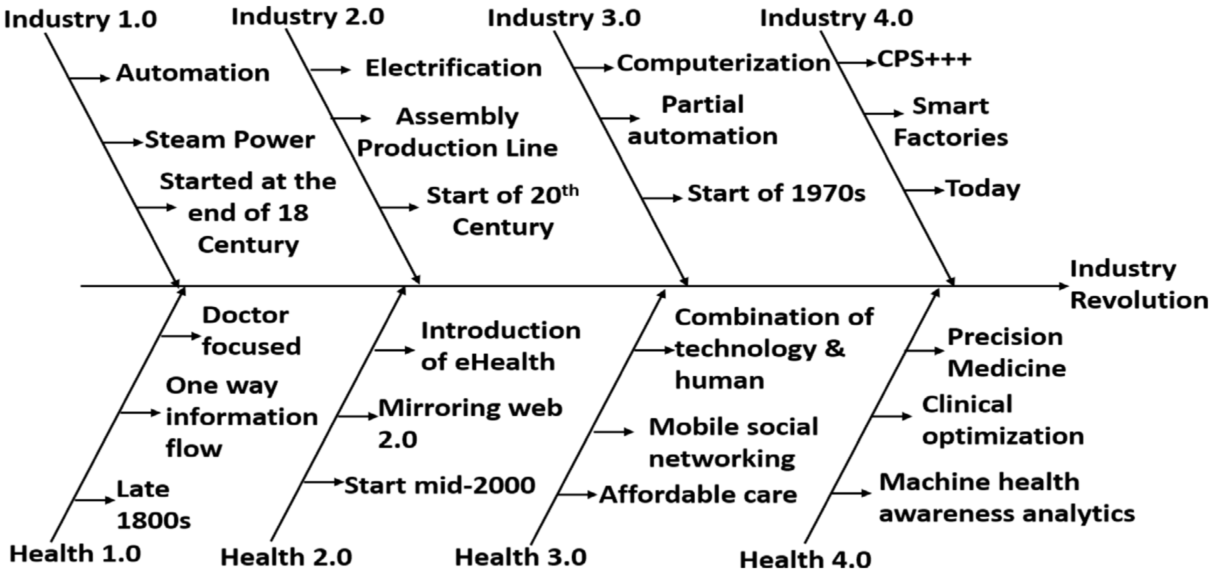


Figure 6.1: All Industrial Revolutions from I1.0 to I4.0 and all health revolutions from H1.0 to H4.0

6.2 Features of I4.0

H4.0 is a subset of I4.0 and came into existence when the development of I4.0 reached some level of maturity. It will become possible after the successful adaptation of the three features of I4.0: (1) vertical integration (2) horizontal integration for the value chain (3) end-to-end engineering for the overall value chain [254] as shown in Fig. 6.2.

Vertical amalgamation requires the digitalization of different hierarchal units of business inside the organization. Vertical amalgamation helps in the transformation of smart factories so that they can generate a more customized product with a great level of profitability. The smart ecosystem can handle the production of different products at a huge level and a high amount of data is processed for operating this manufacturing process easily. This high production rate of the manufacturing process is able to generate different medical instruments and devices that are required according to the patient needs. The integration of horizontal and vertical units of business enables efficiency in the resource allocation process, real-time data sharing, accurate planning, and coherent business units that are crucial for connected devices from the point of view of I4.0.

Horizontal integration is responsible for handling the entire product life cycle by using various ways of financial management, better information system, and material flow [255]. All of them are very important in the health scenario because we are dealing with an abundant amount of data. The extraction of the right information on time is very important in healthcare to take timely actions. Along with this, the development of smart and intelligent health-related products and materials is also crucial for fulfilling the current medical needs of doctors, patients, and sellers. In this product development phase, digital integration work towards end-to-end engineering for product development [256]. This integration is also used for reusing the consistent and continuous product model.

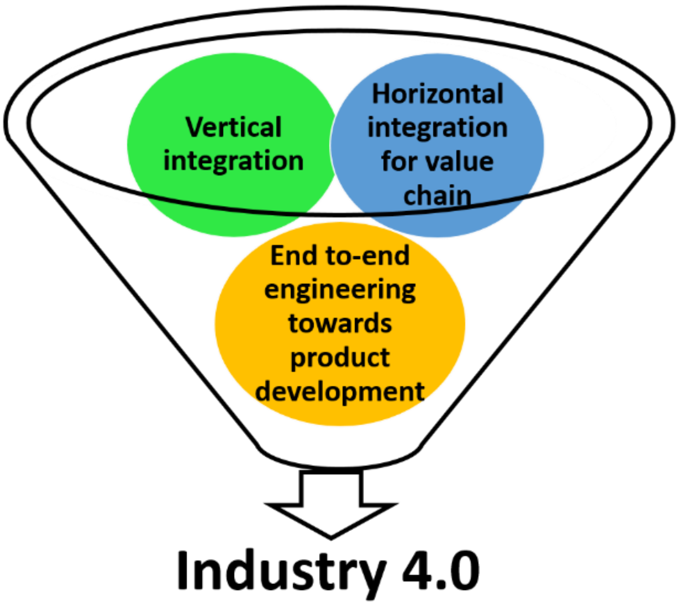

Vertical integration
Horizontal integration for value chain
End to-end engineering towards product development
Industry 4.0

Figure 6.2: Features of I4.0 as vertical integration for the Self-Organized System, Horizontal Integration for An Efficient Ecosystem Among Various Organizations to Share Information and End-to-End Integration Enables Reusability of the Continuous Product Model

6.3 Supportive Technologies for H4.0 via I4.0

H4.0 and I4.0 both are interdisciplinary approach as shown in Fig. 6.3, and suffers from various issues like integration and interoperability among the technologies used in production. To overcome this problem the three-dimensional map of RAMI 4.0 provides various ways in a structured manner to tackle the issues of I4.0 as well as H4.0. It makes sure that all the involved participants must understand each other.

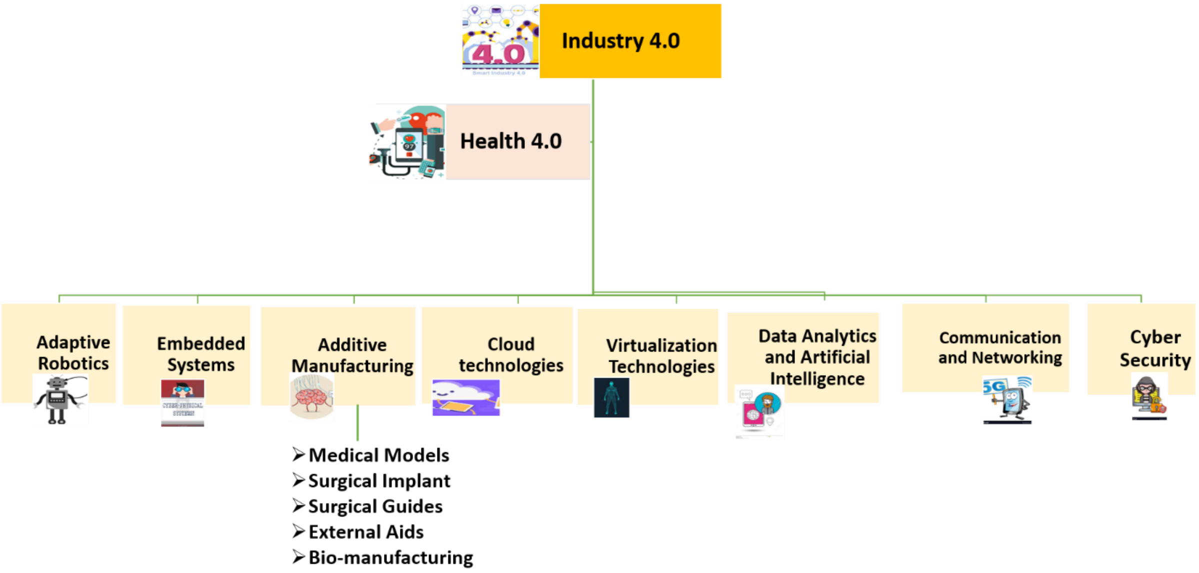


Figure 6.3: Different Technologies That Support The Healthcare Industry To Achieve The Goal of H4.0

6.3.1 Adaptive Robotics

The combination of artificial intelligence with flexible and adaptive robots is used to produce smart products for healthcare. This helps in cost deduction, reducing the waiting time of operations and production time [257]. This advancement in the production field also benefits the health sector as different health equipment is required to monitor the health of the people of different age groups. Robots themselves work as the helping hand of surgeons at the time of surgeries.

6.3.2 Embedded Systems

The embedded system works as a supportive technology for an organization's networking system where it works between the computational capabilities and physical infrastructure [258]. They provide various benefits to the organization such as decision making, intelligent data processing, high-level networking for providing real-time data processing, and feedback from the digital platform.

6.3.3 Additive Manufacturing (AM)

AM technology is used to provide customization according to the patient's requirements. A three-dimensional(3D) sectioned model for an individual patient is developed through customized software [259]. Different models are used in the medical field for obtaining patient data such as Magnetic Resonances Imaging (MRI) technology, Ultrasound, laser scanning, computerized Tomography, etc. [260]. The obtained data shows the different regions of a patient with a similar brightness area and this shows the sections of similar material or particular tissue types. The processing of models is done by using 3-D Computer-Aided Design (CAD) for incorporating various objects, such as fixation devices, implants, etc. After the completion of the virtual model translation of data is performed in Standard Triangulate (STL) format so that it can be used for Rapid Prototyping which is performed through machines [261]. AM is classified into five major areas by Tuomi [262] Medical models, Surgical implant, Surgical guides, External aids, and Bio-manufacturing. It is credible to use AM technology for creating patient's customized or ideal fit implants that help in saving time and money [263, 264, 265].

6.3.4. Cloud Technologies

The integration of cloud technology with I4.0 includes two components. The first component is cloud manufacturing and the second one is cloud computing. Cloud manufacturing can provide a coordinated and linked environment for manufacturing. It is also important to control the supply that is ''available on-demand''. This terminology is used for efficiency enhancement as it helps to reduce the product life cycle costs. This will increase the capability of the service of production systems for making data-driven decisions [266]. Cloud computing is very important for real-time data analysis so that the right information will be delivered on time.

6.3.5. Virtualization Technologies

High-precision surgeries are the main target of the healthcare sector for saving the life of patients and this is possible with the help of Augmented reality and Virtual reality (VR). Augmented reality helps to project the body organs in 3D format for highlighting the problematic zone so that better treatment decisions can be taken. AR and VR technology helps hospitals successfully address simulation-based training, for example, AR technology is used across the world for the application of surgical visualization and vein visualization. AccuVein is used in hospitals for making the procedure easier for healthcare workers to find the vein first time by projecting a map of the patient's veins on their skin.

6.3.6. Data Analytics and Artificial Intelligence

Building a successful big data infrastructure with I4.0 happens only after considering three functions.

- Big data acquisition and integration
- Big data processing and storage
- Big data analytics and knowledge discovery in database.

The first phase includes data collection from smart sensors, Radio Frequency Identification (RFID) readers and RFID tags, etc. RFID tags are used in hospitals to track patients and staff members. The most used technology is passive RFID for the purpose to verify the patient's information. Some hospitals are also using active RFID for tracking the movement of staff and patients. Big data storage and big data processing can be done on two basis. The first one is real-time data processing and the second one is offline processing. Big data analytics and knowledge discovery can be performed by clustering, association, classification, and prediction. To provide precise prediction and classification data must be used in a huge amount so that it can be trained in large quantities and able to do good data preparation so that exact information can be extracted with minimum error rate.

6.3.7. Communication and Networking

Communication and networking work as a link between physically distributed systems. In healthcare different individuals are involved with a patient to provide them better care and they need to share patient information on different platforms [267]. To fulfill this requirement there is a huge need for communication technology to support good health services. Communication and networking are inevitably involved in achieving a better-connected healthcare system. It can be said that information works as blood in healthcare and communication systems works as a heart that pumps the blood. The number of possible conversations is depending on the members involved in a communication channel and the size of the conversation increases as team members increased. It is because the number of possible conversations is calculated by a combinatorial equation where;

$$\text{number of conversations} = n!/(r!(n-r)! \qquad (6.1)$$

Here in Equation, 6.1 n is the number of individuals, and r is the number of individuals involved in a single conversation. Figure 6.4(A) visualized the communication path between three entities and Fig. 6.4(B) shows the increment in the communication path as the number of members increased.

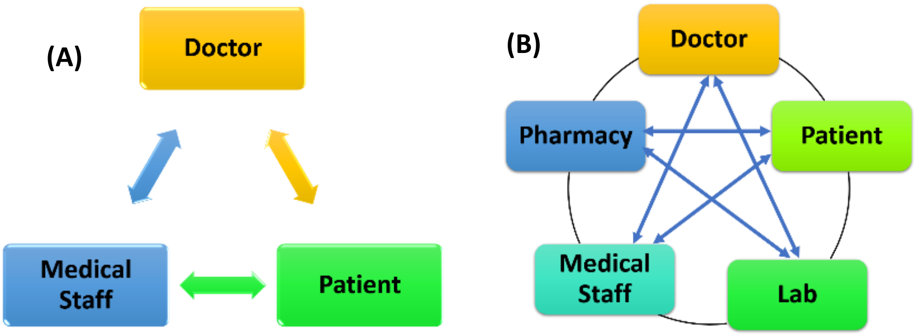


Figure 6.4: (A) Image Depicts the Number of Communication Paths Between The Individuals or Doctor, Patient, Staff (B) Communication Channel Increases As The Number of Individuals Increase In This Channel Who Needs To Communicate

6.3.8. Cyber Security

The collaboration between healthcare and I4.0 requires excessive data gathering and data processing steps. The healthcare network is a cyber-physical system that contains many interconnected physical systems and these systems use cyber technology for interaction [155]. This interconnection of different networking devices introduces a vulnerability from security point of view. Now, providing security of data becomes the fundamental requirement. Four specific threats are manifested in healthcare are:

- data loss
- monetary theft
- attacks on medical devices
- attack on infrastructure

6.4 Design Principles of I4.0 to Implement in the Health Sector

The six design principles of I4.0 to implement H4.0 are as follows:

6.4.1 Interoperability

H4.0 requires a dynamic optimization process for production lines and delivering services. It is a combination of heterogeneous devices that works together in a cooperative way to achieve a common goal. After doing lots of investigation it was found that interoperability was highlighted rapidly in IoT discussion and also discussed health in the white paper of EU-China [268]. In ‘‘smart factories’’ and the ‘‘future factories’’ interoperability plays an important role in communication between all levels of manufacturing, cloud systems, AI, and data analytics. They all have a major contribution in respect to the changing requirements of system stability.

6.4.2 Virtualization

In the medical field, virtualization plays an important role and it is required in different stages to virtualize the physical process. The process of virtualization helps in bringing the healthcare services closer to the patients by providing easy communication between caregivers through digital technology that can be run on laptops or desktops, phones, tablets and other customized devices.

6.4.3 Real-Time Capability

In healthcare real-time capability is like an orchestration process where things are interconnected to achieve the target on time. As in the current scenario where the whole world suffers from the COVID-19 problem and hospitals are overburdened. Patients get their treatment irrespective of their physical presence inside or outside of the hospital in the exact amount of medication required for increasing therapeutic effects and reducing side effects. This process leads to an improved healthcare environment without increasing the burden on limited physical health infrastructure.

6.4.4 Decentralization

The new future of the healthcare sector is a distributed or decentralized patient-centred model. The distributed environment of a patient-centred model requires information flow across various domains and networks.

6.4.5 Modularity

Modular systems are used because they are flexible to adapt, change if required, or expand any individual module. The advantage of this system is explained through a European commission FI-STAR project for the health domain. Under this project software components like specific Enabler and Generic Enabler were utilized for creating new functionality of recombining various activity groups.

6.4.6 Service Orientation

Service aggregation is acknowledged as a way to provide customer-centred services and therefore becomes an integral part of the H4.0 design principle. Nowadays pharmaceutical companies take a clear shift from playing the role of drug manufacturing to a service provider. This all happens after harvesting the information gathered from big data of different sensors such as smart inhalers data so that a better facility is provided by them. Pharma companies play a dual role from a business model point of view first is selling the drugs and the second is providing a service for the management of diseases.

6.5 Technology Roadmap to Achieve H4.0 by Implying I4.0

To achieve H4.0 by implying I4.0 as a digital enterprise it is required to build a technology roadmap accurately. The technology roadmap is a vital method and acts as an integral part of producing and delivering an innovative strategy to many organizations. The graphical nature of the roadmap helps in strategic alignment. The roadmap technology helps in addressing the business performance objectives of a company to achieve success in market position [269]. The architecture presented by Phaal et al. in 2001 was used as a tool for firms’ policymakers and managers for building a roadmap. The work presented by them becomes a primary framework for the technology-push approach and market pull approach. The drawback of this architecture is that the enterprises are unable to launch roadmaps due to the scarcity of qualified staff. This drawback is addressed in [271] by concluding that qualified specialist handles it in long-term planning for the road mapping process. This strategy is efficient in many ways like it tells the benefit of time management and tells the usage of setting competitive targets. In [272] Phaal et al. presented eight types of graphical roadmaps based on specific needs, these recursive tools are useful for strategic management. In a study presented by Cancer et al. in 2007 [270] the most useful technology roadmap for current and future states relates to three primary levels such as market, product, and resources multiple-layer type. In 2016 a study presents a roadmap where R&D contributes toward technological factors while marketing shows a product-related viewpoint [271, 272, 273]. This roadmap helps in building a competitive strategy. All of these techniques and their explanation are shown in Table 6.1. The benefits and challenges of roadmap technology are presented in Table 6.2.

A ‘‘roadmap’’ is used to empower the management of the industry to clearly understand the moves and decisions to be taken. This procedure is explained in Fig. 6.5 where the plan and the key technologies were explained in the first phase and subsequently, the second phase contains the development of a new product.

Strategy: A strategy is a time-based plan that portrays the current position of the industry, the future where it should be, and the way to achieve it. In the roadmap, the strategy phase is a collaborative way of planning. The three-dimensional map of RAMI 4.0 helps to achieve this objective by showing all the issues of I4.0 in a structured way. To prepare the H4.0 roadmap the first step should be estimating the digital maturity of the enterprise for setting the targets based on time for the next years. This step helps to think about the future of the organization and in the case of H4.0 this transformation helps not to think about the ‘if’ for this industry instead of this it is a matter of ‘when’. This transformation is tough to predict but it will be going to a transformation from hospital-centred healthcare to distributed patient-centred healthcare to self-administered healthcare. The roadmap should consider the changes that happen in the future concerning the patient’s behaviour and how the relationship of the organization will change with them.

Product Development Phase: The new product development phase is used to sketch goals by considering different constraints and principles. Constraints can be considered in distinct ways, in this study we considered technology constraints and goals. Technology constraints are important from a budget point of view and a partnership point of view from other industries. H4.0 requires the collaboration of different technologies and organizations and all of this should require a certain budget that decides the saving and revenues of the health industry. This thing starts from concept generation then the assessment step is performed after this the potential of the project is calculated and based on the potential one of the projects will get selected. Then finally prioritise it according to the potential and implement it. In the current scenario, the ongoing research on the SARS-CoV-2 virus doctors and scientists are changing their line of treatment according to the symptoms of the virus. This helps them to apply the same line of treatment to the mass population so that they will get the maximum advantage from this research. These steps are performed in advance to increase the success rate.

Table 6.1: Year-wise Roadmap Techniques Along with their Explanation

S.no	Techniques	Explanation	Year	References
1	T-plan	It gives a framework for firms based on the market-pull strategy that tells the use of road mapping by utilizing minimum resources.	2001	270
2	(a)	Improved communication across the team by identifying communication gaps and technology gaps.	2004	271
3	(b)	Eight types of graphical roadmaps are presented along with the usefulness of strategic management.	2004	272
4	Driving factors for the R&D	R&D and marketing make a conjoint roadmap.	2016	273

Table 6.2: Benefits and Challenges of Roadmap Techniques

S.No	Benefits of Technology Roadmaps	Challenges of Technology Roadmaps
1	Building an alignment of technical and commercial strategies	Lesser knowledge of technology and opportunities related to them.
2	Helps in setting rational and competitive targets	Lesser knowledge about new product demands of customers and business models comes under the vision of I4.0.
3	Gives a clear view to prioritize the investment	Limited resources in terms of finance and manpower.

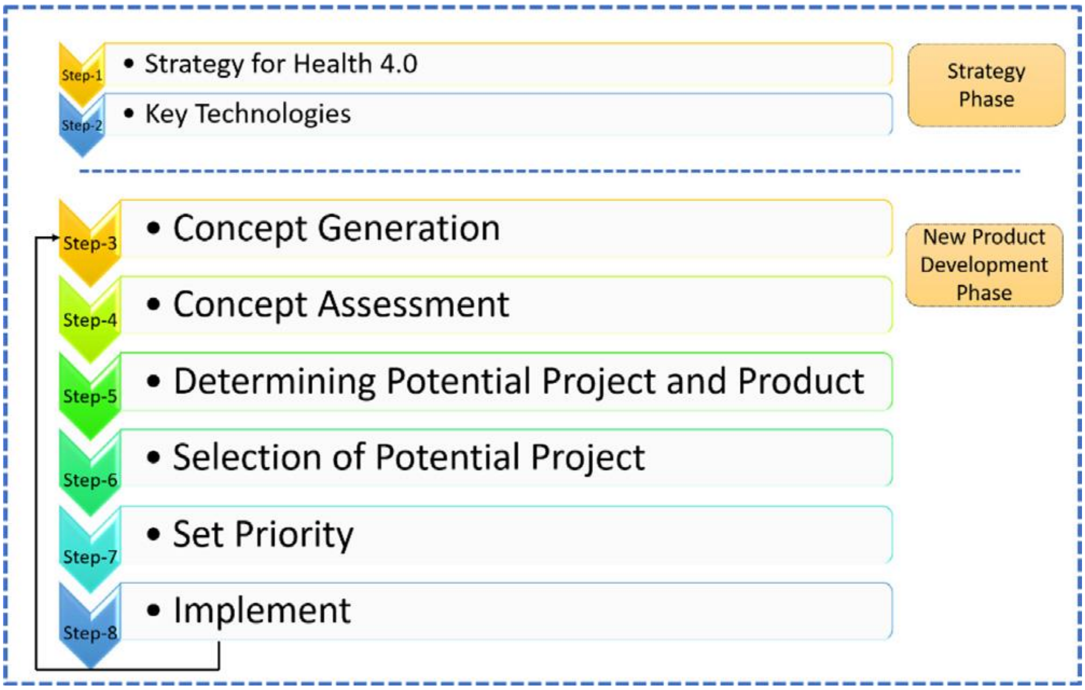


Figure 6.5: Proposed Technology Roadmap for H4.0

6.6 Smart eHealth Architecture of H4.0 by Implying I4.0

The proposed eHealth architecture consists of four layers and it is a complete solution of health 4.0 via I4.0. This architecture combines all the healthcare-related individual entities like clinics, hospitals, policymakers, smart health devices production companies, government officials, etc to work in coordination for providing a patient-centred environment. As depicted in Fig. 6.6 the architecture contains four layers data namely, data collection layer, data access layer, data implementation layer, and data application layer.

6.6.1 Data Collection Layer

Information and communication technology helps in providing the availability of anytime-and-anywhere connectivity. It is further required by every sector of I4.0 and H4.0 to communicate throughout the hierarchical level. This unprecedented ubiquitous presence of mobile and wireless technologies also empowers the healthcare sector of developing countries. The accessibility of cost-efficient services and low-cost miniaturized wireless sensors have enabled new cost-efficient services for healthcare. These services are the main cause of healthcare data generated from various devices such as smart homes, health devices, IoT data, robot data who is working as help to patients and surgeons, etc.

The data collection layer of the proposed architecture first collects the data and started processing it locally on a fog node to reduce cloud load. The fog computing helps in significantly mitigating various issues faced by the cloud while information is transferred such as delay, cost overhead, and jitter. Fog computing helps in improving on-time service delivery. Moreover, fog computing worked as a powerful tool for intelligent processing of decentralized unprecedented data volumes produced by IoT sensors deployed to integrate cyber and physical environments that helps the IoT to achieve its vast potential. In this layer, a mechanism known as RTM (Real Time Monitoring) provided by fog is also used for emergencies and the direct call is given to the hospital, associated family members, friends, and ambulance. Then finally the data which is not required for real-time processing is sent to the eHealth cloud.

6.6.2 Data Access Layer

The architecture supports the collection of health-related data for the remote monitoring of patients through a number of sensors and applications. The advanced environmental and medical sensors integrated into a mobile phone and personal health devices are used for sending information like humidity and temperature, ECG, blood pressure, glucose, etc. and they also continuously monitor patients’ physical and physiological conditions. In the data access layer hospitals used this data for taking the health decisions of the patients, so that doctors can assist patients even if they are not physically present or admitted to the hospital. This facility reduces hospital load by handling more patients. Hospitals maintain their information system by using fog computing and with the help of sensor bands they can easily monitor their staff and patients. This smart technology helps the hospital authorities to provide personalized treatment to the patients. The delivery of customized health equipment becomes an easy task as the understanding of patients’ requirements is more clear and precise than earlier. The technology also supports the authorities to maintain the record of medicines they require in the future and transfer the information to the next level. This entire data will be stored in the cloud so that related BDA studies can be performed for further analysis. Some of the recent studies have highlighted the significance of H4.0. In the study of Aceto et al. [1], the authors have mentioned the importance of medication intake monitoring with the help of RFIDs, Sensors and mobile apps. Even some of the advanced methods provide the Proteus digital health. It gives an insight view of patient health and helps the practitioners to take the right decision about medication treatment. Another practical implementation of H4.0 has been seen in the case of asthma care for critically ill patients where smart inhalers help patients as well as doctors to maintain the record of proper dosage. Chute and French [34] have stated the benefit of H4.0 in reducing transactional costs while improving treatment quality and service. The authors provided a proper toolset of I4.0 for asthma care by mentioning the advantages of the virtual agent in case of personalised care for doctors and patients based on unknown and known triggers. These triggers provide information about regional air pollution and offer connectivity to users from other communities for broader support.

6.6.3 Data Implementation Layer

To provide personalized healthcare, smart factories and smart pharmaceuticals play a vital role. The smart factories produce their products based on the specification of patients and take patients' specific decisions. The smart factories and smart pharmaceuticals deliver smart products that use implantable micro and nano technologies such as implantable insulin pumps, fall detectors, defibrillators vests, etc. They try to understand the biology of individuals whose data is present to impact diagnosis, predisposition, screening, pharmacogenomics, prognosis and surveillance. To fulfil all the requirements future pharmaceuticals are likely to collect macro and micro-level metadata that may offer new insights into disease and provide personalized healthcare.

In the data implementation layer of the eHealth architecture, smart factories and smart pharmaceuticals are directly taking their orders from the hospitals. Smart management takes smart decisions to achieve more benefits from the planning stage to the production stage and finally the packaging stage. To improve their productivity in the implementation layer smart factories are using dew computing. Dew computing is used to increase the efficiency of the factories and take full advantage of technology.

6.6.4 Data Application Layer

Healthcare data is present in various formats and scattered forms. Different organizations have their rights on the data which increases the difficulty level to manage the data in a central place. As a solution, this layer came into existence and it proposes to maintain all healthcare data on a separate eHealth cloud. The administrative control is given to the government officials along with that access is provided to individuals and different organizations to whom the data belongs. This layer is used by policymakers to formulate better policies after considering all parameters. Policymakers have full control of the eHealth cloud to use the data for research, marketing decisions, and the making of laws. The architecture present in Fig. 6.6 provides a complete solution from data collection, data access, data implementation, and data application in a cloud platform where fog and dew computing is used to decrease the load of the eHealth cloud.

In India, the population density is very high and it has the world's second-largest internet user population. This digital journey of the country gives exuberant power to the government and other supporting organizations to use this information for creating new opportunities in different areas. H4.0 is the solution to provide better services to the vast range of the population by connecting them with the internet and it will be very helpful to deal with pandemic situations easily. As it is seen in this COVID-19 period government officials are using the internet to provide information regarding containment zones, COVID hospitals, availability of a bed, etc., through specific mobile applications.

The official application used by the Indian government is Aarogya Setu to provide information to the users about their potential risk of getting infected from this Covid-19 infection. The Aarogya Setu was launched on 02 April 2020 to provide information about the spread of the Corona Virus throughout the country. Meanwhile, some of the state governments are also using some different applications for the awareness of their local citizens like the Delhi government launches Delhi Corona application. This application provides information regarding Covid-19 hospital's status in Delhi. It works without registration and the home screen of the application tells about the total number of Covid-19 beds and how many of them are occupied and vacant. It gives clear information about the ICU total beds with and without ventilators and also tells about the non-Covid-19 ICU beds. This type of clear information helps the patients as well as doctors to make the right decision at the time of any medical emergency. Another very important application launched by the Indian government for Covid-19 vaccination is the Co-Win app. This application also works on various modules for different users based on their roles. All the Covid-19 vaccine-related necessary data for proper vaccination in the country is uploaded on this application. This digital information provides transparency of the available resources so that officials can take proactive decisions based on the conditions and patients can also approach the right place without any wastage of time.

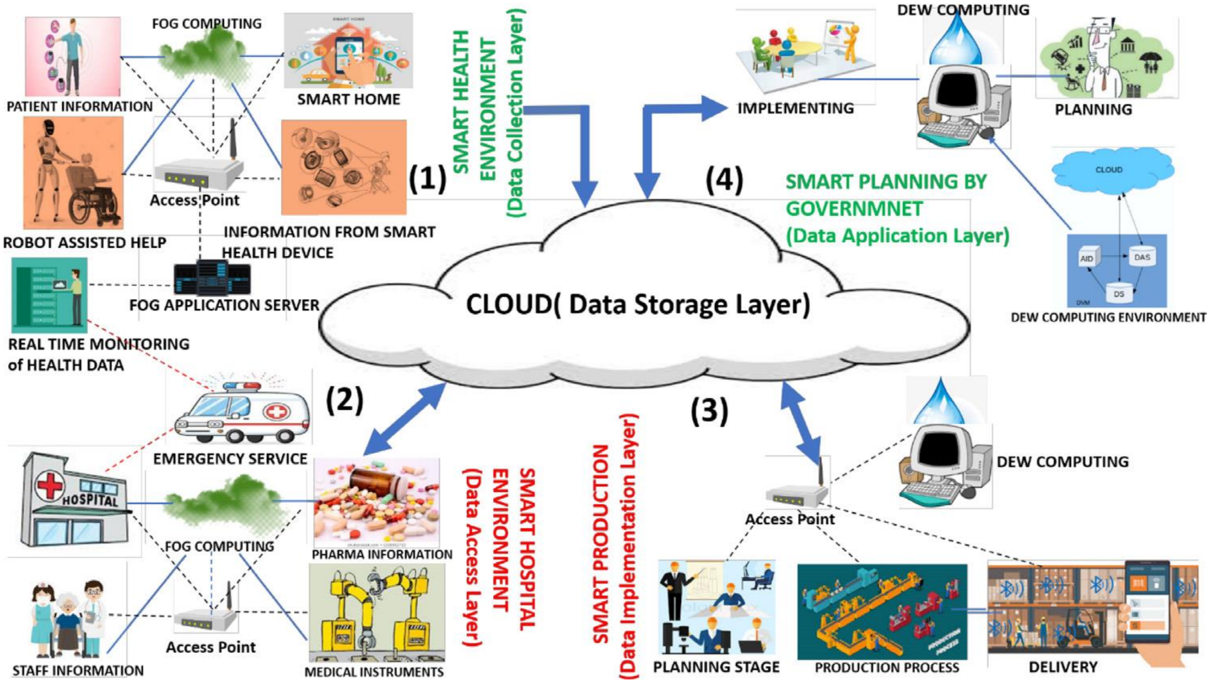


Figure 6.6: The Proposed eHealth Architecture for obtaining H4.0

6.7 Summary

This chapter presents an eHealth architecture along with the roadmap to achieve the target of H4.0. The chapter also presents the design principles of I4.0 to implement in the health sector. The proposed architecture contains four layers for different purposes. The data collection layer performs the task of data collection and the data access layer utilizes the data to provide better healthcare facilities and also utilizes it for personal management.

After this, the data implementation layer utilizes all the collected data to provide personalized healthcare. Moreover, a data application layer offers cloud storage of healthcare data for policymakers. The policymakers have full control of the cloud data to take decisions and make laws.

CHAPTER 7

CONCLUSION AND FUTURE WORK

This chapter summarizes the work done in the thesis, presents contributions, and highlights the future directions.

This research work presents a study for healthcare, where initially it highlights the impact of I4.0 on H4.0, explains the benefits of RAMI 4.0, and presents the different axis of healthcare big data where it contains eight Ps of healthcare related things. Then it focuses on various environmental diseases and discuss various causes of the environmental diseases. Later on, it presents an HCMP architecture for healthcare that works on parallel and distributed Hadoop architecture. The proposed architecture provides an improved data storage and management scheme for healthcare data. Along with this an algorithm (MySymptom) has been proposed and utilised for the selection of CKD patients based on their symptoms. Further, an improved OHCMDD architecture was framed and presented for multiple disease detection with a hybrid machine learning model for disease prediction.

Moreover, in addition to the proposed architectures the present study also proposed an H4.0 architecture to handle different parameter of the healthcare industry. It works on the design principles of the I4.0. The performance of the proposed architecture has compared with the state-of-the-art techniques and has given significant results for different parameters. The experimental results, performance analysis, and comparative analysis demonstrate the effectiveness and efficiency of the proposed system.

7.1 Summary of Work Done in the Thesis

This section summarizes the thesis to resolve the challenges and issues present in the existing healthcare system.

1. To address the shortcomings of the existing healthcare architectures this research work proposed a novel healthcare architecture that works on different layers for providing better healthcare services.
2. The proposed architecture works on Hadoop ecosystem for proper storage and maintenance of the healthcare data. For this healthcare data management two algorithms have been proposed that works for providing robust data storage technique.
3. Further for better understanding the scaling has been performed on the Hadoop cluster where DataNode are added in the existing cluster.
4. To eliminate the limitations of the HCMP architecture a new OHCMDD architecture has been proposed and implemented where it works on multiple diseases with hybrid D-SMO algorithm and DataNode decommissioning problem.
5. To optimize the overall healthcare industry a new H4.0 architecture was proposed. It works on the design principles of the I4.0. It also offers the road map to achieve the target of H4.0.

7.2 Contributions of the Research

This research work describes the solution approach utilized to address the challenges of healthcare organisation. We proposed the healthcare architecture based on distributed scheme for better storage, management and analysis of data to provide better facilities for overall healthcare industry and patients; comprising of following features:

- **Healthcare data management using data distribution algorithm**
- **Disease prediction based on the symptoms**
- **Optimisation of healthcare system**
- **Provides H4.0 architecture for overall healthcare industry**

1. This work proposed a HCMP architecture for healthcare. The proposed architecture implements the data distribution algorithms that works on Hadoop ecosystem that work for the proper management of the healthcare data. The algorithms works on the heterogenous Hadoop cluster where they distributes the data among the DataNodes based on their CR. The DataNode who finishes its task gets the more number of jobs in comparison to the slow DataNode. The data handling by the proposed data distribution algorithms was presented by considering three cases where a new job type arrives in

the existing cluster, different job type arrives in the existing cluster, and finally old job type arrives in the cluster where a new DataNode was admitted.

2. The CKD disease was predicted based on the symptoms pool and accordingly the scores were assigned to the patients. The patients whose score is near to one may be affected by the CKD disease and the patients whose score is below the threshold value is not affected by CKD.
3. The proposed work also presents a new healthcare architecture that works on the limitations of the previous architecture to provide more robustness and elasticity by handling multiple diseases with a new D-SMO algorithm for better prediction.
4. The proposed architecture further enhanced to achieve the target of H4.0 by implementing the design principles of I4.0. This research also presents an architecture for H4.0 to improve the overall healthcare organisation.

7.3 Future Work

We plan to deploy the proposed architecture in the form of eHealth cloud that works on different parameters for healthcare industry to provide reliable, dynamic, and safe management of healthcare related things.

For future analysis work, inverse frequency and vector similarity may be used to search the less common symptoms of the disease dataset through MySymptom algorithm. This will help to study the effect of these symptoms on any other chronic disease.

Publications Related to the Thesis

Papers Published in International Journals:

- Sisodia, Amrita, and Rajni Jindal. "An effective model for healthcare to process chronic kidney disease using big data processing." *Journal of Ambient Intelligence and Humanized Computing* (2022): 1-17. **(SCIE, IF: 7.104)**
- Sisodia, Amrita, and Rajni Jindal. "A meta-analysis of industry 4.0 design principles applied in the health sector." *Engineering Applications of Artificial Intelligence*, 104 (2021): 104377. **(SCIE, IF:6.212)**

Papers Communicated in International Journal

- Amrita Sisodia and Rajni Jindal. "Investigation of multi-disease diagnostic system for healthcare using hybrid D-SMO algorithm" Springer Journal. (Communicated)
- Amrita Sisodia and Rajni Jindal. "Improving the Hadoop ecosystem using YARN for healthcare services" Elsevier. (Communicated)

Papers Published in International Conferences:

- Sisodia, Amrita, and Rajni Jindal. "Exploring the application of big data analysis in healthcare sector." In *2017 International Conference on Computational Science and Computational Intelligence (CSCI)*, pp. 1455-1458. IEEE, 2017.
- Sisodia, Amrita, and Rajni Jindal. "Big Data Analysis for Prediction of Environmental Diseases." 2019 Sustainable Technologies for Environmental Management (STEM). Scopus,2019, DTU.
- Sisodia, Amrita, and Rajni Jindal. "Prediction of Environmental Diseases Using Machine Learning." In *International Conference on Innovative Computing and Communications*, pp. 521-531. Springer, Singapore, 2022.
- Sisodia, Amrita, and Rajni Jindal. "Analysis of the Corona Virus Outbreak Using Machine Learning." In 2021 6th International Conference on Advance Production and Industrial Engineering (ICAPAIE).

Published Book Chapters:

- Sisodia, Amrita, and Rajni Jindal. "Prediction of Environmental Diseases Using Machine Learning." In *International Conference on Innovative Computing and Communications*, pp. 521-531. Springer, Singapore, 2022.

References

[1] G. Aceto, V. Persico, and A. Pescapé, "The role of Information and Communication Technologies in healthcare: taxonomies, perspectives, and challenges," Journal of Network and Computer Applications, vol.107, pp. 125-154, 2018.

[2] F.A. Alaba, M. Othman, I.A.T. Hashem, F. Alotaibi, “Internet of things security: a survey,” Journal of Network and Computer Applications, vol. 88, pp. 10–28, 2017.

[3] C. Pino, R. Di Salvo, “A survey of cloud computing architecture and applications in health,” In: International Conference on Computer Science and Electronics Engineering (ICCSEE), 2013, pp. 1649-1653.

[4] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, AH. Byers, “Big data: The next frontier for innovation, competition, and productivity,” McKinsey Global Institute, 2011, Accessed 17 Jan 2019.

[5] D. Ramesh, P. Suraj, L. Saini, “Big data analytics in healthcare: a survey approach,” In: International Conference on Microelectronics, Computing and Communications (MicroCom), IEEE, 2016, pp. 1–6.

[6] S. Shafqat, S. Kishwer, R. U. Rasool, J. Qadir, T. Amjad, and H. F. Ahmad, “Big data analytics enhanced healthcare systems: a review,” The Journal of Supercomputing, vol. 76(3), pp.1754-1799, 2020.

[7] G. Riva, “Ambient intelligence in health care,” Cyber Psychology & Behavior, vol. 6(3), pp. 295–300, 2014.

[8] J. T. Pollettini, S. R. G. Panico, J. C. Daneluzzi, R. Tinós, J. A. Baranauskas, A. A. Macedo, “Using machine learning classifiers to assist healthcare-related decisions: classification of electronic patient records,” Journal of medical systems, vol. 36 (6), pp. 3861–3874, 2012.

[9] M. M. Ahsan, and Z. Siddique, “Industry 4.0 in Healthcare: A systematic review,” International Journal of Information Management Data Insights, vol. 2(1), pp.1-14, 2022.

[10] Z. A. Abas, Z. Z. Abiding, A. F. N. A. Rahman, H. Rahmalan, G. Pramudy, and M. H. A. Hamid, “Internet of Things and Healthcare Analytics for Better Healthcare Solution: Applications and Challenges,” International Journal of Advanced Computer Science and Applications, vol. 9(9), pp. 446-450, 2018.

[11] S. H. Almotiri, M. A. Khan, and M. A. Alghamdi, “Mobile health (m-health) system in the context of IoT,” In: IEEE 4th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW), 2016, pp. 39-42.

[12] M. N. Ahmed, A. S. Toor, K. O'Neil, and D. Friedland, “Cognitive computing and the future of health care cognitive computing and the future of healthcare: the cognitive power of IBM Watson has the potential to transform global personalized medicine,” IEEE pulse, vol. 8(3), pp. 4-9, 2017.

[13] G. Aruni, G. Amit, and P. Dasgupta, "New surgical robots on the horizon and the potential role of artificial intelligence" Investigative and clinical urology, vol. 59(4), pp. 221- 222, 2018.

[14] Y. Bhatt and C. Bhatt, "Internet of things in healthcare," In: Internet of things and big data technologies for next generation HealthCare, Springer, 2017, pp. 13-33.

[15] C. Patrone, M. Lattuada, G. Galli, and R. Revetria, "The role of internet of things and digital twin in healthcare digitalization process," In: The World Congress on Engineering and Computer Science, Springer, 2018, pp. 30-37.

[16] P. D. Kaur and I. Chana, "Cloud based intelligent system for delivering health care as a service," Computer methods and programs in biomedicine, vol. 113(1), pp. 346-359, 2014.

[17] T. N. Gia, M. Jiang, A.-M. Rahmani, T. Westerlund, P. Liljeberg, and H. Tenhunen, "Fog computing in healthcare internet of things: A case study on ecg feature extraction," In: Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing (CIT/IUCC/DASC/PICOM), 2015, IEEE International Conference, pp. 356-363.

[18] M. M. Mahmoud, J. J. Rodrigues, S. H. Ahmed, S. C. Shah, J. F. AI-Muhtadi, V. V. Korotaev and V. H. C. Albuquerque, "Enabling technologies on cloud of things for smart healthcare," IEEE Access, vol. 6, pp. 31950–31967, 2018.

[19] A. Ustundag, and E. Cevikcan, "Industry 4.0: managing the digital transformation," Springer, 2017.

[20] L. Atzori, A. Iera, and G. Morabito, "The internet of things: A survey," Computer Networks, vol. 54(15), pp. 2787-2805, 2010.

[21] J. Lee, "Smart health: Concepts and status of ubiquitous health with smartphone," In: ICTC, IEEE, 2011, pp. 388-389.

[22] J. Gantz and D. Reinsel, "Extracting value from chaos," Technical Report IDC iview 1142, pp. 1-12, 2011.

[23] F. Martin-Sanchez and K. Verspoor, "Big data in medicine is driving big changes," Yearbook of medical informatics, vol. 23(01), pp. 14-20, 2014.

[24] L. Rajabion, A. A. Shaltooki, M. Taghikhah, A. Ghasemi, and A. Badfar, "Healthcare big data processing mechanisms: The role of cloud computing," International Journal of Information Management, vol. 49, pp. 271-289, 2019.

[25] P. A. Laplante and N. L. Laplante, "A structured approach for describing healthcare applications for the internet of things," In: Internet of Things (WF-IoT), IEEE 2nd World Forum on. IEEE, 2015, pp. 621-625.

[26] P. A. Laplante and N. Laplante, "The internet of things in healthcare: Potential applications and challenges," IT Professional, vol. 18(3), pp. 2-4, 2016.

[27] M. Wehde, "Healthcare 4.0.," IEEE Engineering Management Review, vol. 47(3), pp. 24-28, 2019.

[28] J. Archenaa and E. M. Anita, "A survey of big data analytics in healthcare and government" Procedia Computer Science, vol. 50, pp. 408-413, 2015.

[29] H. Cao, V. Leung, C. Chow, and H. Chan, "Enabling technologies for wireless body area networks: A survey and outlook," IEEE Communications Magazine, vol. 47 (12), pp. 84-93, 2009

[30] M. Chen, S. Gonzalez, A. Vasilakos, H. Cao, and V. C. Leung, "Body area networks: A survey," Mobile networks and applications, vol. 16(2), pp. 171-193, 2011.

[31] A. J. Trappey, C. V. Trappey, U. H. Govindarajan, A. C. Chuang, and J. J. Sun, "A review of essential standards and patent landscapes for the internet of things: A key enabler for industry 4.0.," Advanced Engineering Informatics, vol. 33, pp.208-229, 2017.

[32] C. L. Ventola, "Mobile devices and apps for health care professionals: uses and benefits," Pharmacy and Therapeutics, vol. 39(5), pp. 356-64, 2014.

[33] R. Chen and M. Snyder, "Promise of personalized omics to precision medicine," Wiley Interdisciplinary Reviews: Systems Biology and Medicine, vol. 5(1), pp. 73-82, 2013.

[34] C. Chute, and T. French, "Introducing care 4.0: an integrated care paradigm built on industry 4.0 capabilities," International journal of environmental research and public health, vol. 16(12), pp. 1-17, 2019

[35] I. Azodo, R. Williams, A. Sheikh, and K. Cresswell, "Opportunities and challenges surrounding the use of data from wearable sensor devices in health care: qualitative interview study," Journal of medical Internet research, vol. 22(10), pp. 1-12, 2020

[36] S. M. R. Islam, D. Kwak, M. H. Kabir, M. Hossain, and K. S. Kwak, "The internet of things for health care: A comprehensive survey" IEEE Access, vol. 3, pp. 678-708, 2015.

[37] C. Doukas and I. Maglogiannis, "Bringing IoT and cloud computing towards pervasive healthcare," In: Sixth International Conference Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2012, IEEE, pp. 922-926.

[38] N. Terry, "Will the internet of things disrupt healthcare," Vand. J. Ent. & Tech. L., vol. 19, pp. 327, 2016.

[39] C. Thuemmler, C. Bai, "Health 4.0: How Virtualization and Big Data Are Revolutionizing Healthcare," Springer: Basel, Switzerland, 2017, pp. 2168-2194.

[40] A. Annunziato, "5G vision: NGMN – 5G initiative," In: IEEE Conference 81st Vehicular Technology (VTC Spring), 2015, pp. 1-5.

[41] W. D. De Mattos and P. R. L. Gondim, "M-health solutions using 5G networks and M2M communications," IT Professional, vol. 18(3), pp. 24-29, 2016.

[42] M. Hermann, T. Pentek, O. Boris, "Design Principles for Industrie 4.0 Scenarios: a literature review," Technical University of Dortmund: Dortmund, pp. 1-15, 2015.

[43] P. Ajmera, and V. Jain, "Modelling the barriers of Health 4.0–the fourth healthcare industrial revolution in India by TISM," Operations Management Research, vol. 12(3), pp.129-145, 2019.

[44] H. B. Jamkhaneh, G. L. Tortorella, S. V. Parkouhi, and R. Shahin, “A comprehensive framework for classification and selection of H4. 0 digital technologies affecting healthcare processes in the grey environment,” The TQM Journal, pp. 1-28, 2022.

[45] Y. L. Zheng, X. R. Ding, C. C. Y. Poon, B. P. L. Lo, H. Zhang, X. L. Zhou, G. Z. Yang, N. Zhao, and Y. T. Zhang, “Unobtrusive sensing and wearable devices for health informatics,” IEEE Transactions on Biomedical Engineering, vol. 61(5), pp.1538-1554, 2014.

[46] K. Kang, Z. B. Pang, L. D. Xu, L. Y. Ma, and C. Wang, “An Interactive trust model for application market of the internet of things,” IEEE Transactions on Industrial Informatics, vol. 10(2), pp. 1516–1526, 2014.

[47] G. Yang, L. Xie, M. Mäntysalo, X. Zhou, Z. Pang, Xu L. Da, S. Kao-Walter, Q. Chen, and L. R. Zheng, “A health-IoT platform based on the integration of intelligent packaging, unobtrusive bio-sensor, and intelligent medicine box” IEEE transactions on industrial informatics, vol. 10(4), pp. 2180-2191, 2014.

[48] Z. Pang, G. Yang, R. Khedri, and Y. T. Zhang, “Introduction to the special section: convergence of automation technology, biomedical engineering, and health informatics toward the healthcare 4.0.” IEEE Reviews in Biomedical Engineering, vol. 11, pp.249-259, 2018.

[49] P. Sobradillo, F. Pozo, Agust, Lvar, “P4 medicine: the future around the corner,” Archivos de Bronconeumologa ((English Edition)), vol. 47(1), pp.35–40, 2011.

[50] A. Sonawane, P. Manickam, and S. Bhansali, “Stability of enzymatic biosensors for wearable applications,” IEEE reviews in biomedical engineering, vol.10, pp. 174–186, 2017.

[51] G. N. Vilaza, and J. E. Bardram, “Sharing Access to Behavioural and Personal Health Data: Designers' Perspectives on Opportunities and Barriers,” In: Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare, 2019, pp. 346-350.

[52] C. Habib, A. Makhoul, R. Darazi, and C. Salim, “Self-adaptive data collection and fusion for health monitoring based on body sensor networks,” IEEE Transactions on Industrial Informatics, vol. 12(6), pp. 2342-2352, 2016.

[53] E. Omanovic-Mikli ´ canin, M. Maksimovi ˇ c, and V. Vujovi ´ c, “The future of ´ healthcare: Nanomedicine and internet of nano things,” Folia Medica Facultatis Medicinae Universitatis Saraeviensis, vol. 50(1), pp, 23-28, 2015.

[54] R. Li, D. T. H. Lai, and W. Lee, “A survey on biofeedback and actuation in wireless body area networks (WBANs),” IEEE reviews in biomedical engineering, vol. 10, pp. 162–173, 2017.

[55] N. Alshurafa, W. Xu, J. J. Liu, M. C. Huang, B. Mortazavi, C. K. Roberts, and M. Sarrafzadeh, “Designing a robust activity recognition framework for health and exergaming using wearable sensors,” IEEE Journal of Biomedical and Health Informatics, 18(5), pp.1636-1646, 2013.

[56] J. E. Bardram, “Pervasive healthcare as a scientific discipline,” Methods of information in medicine, vol. 47(03), pp.178-185, 2008.

[57] X. Shen, J. Misic, N. Kato, P. Langenörfer, and X. Lin, “Emerging technologies and applications of wireless communication in healthcare” Journal of Communications and Networks, vol. 13(2), pp. 81-85, 2011.

[58] G. Huzooree, K. Kumar Khedo, and N. Joonas, “Pervasive mobile healthcare systems for chronic disease monitoring,” Health informatics journal, vol. 25(2), pp. 267-291, 2019.

[59] G. S. Gunanidhi, “Extensive analysis of Internet of Things based health care surveillance system using Rfid assisted lightweight cryptographic methodology,” Turkish Journal of Computer and Mathematics Education, vol. 12(10), pp. 6391-6398, 2021

[60] M. Haghi, S. Neubert, A. Geissler, H. Fleischer, N. Stoll, R. Stoll, and K. Thurow, “A flexible and pervasive IoT-based healthcare platform for physiological and environmental parameters monitoring,” IEEE Internet of Things Journal, vol. 7(6), pp. 5628-5647, 2020.

[61] S. Hiremath, G. Yang, and K. Mankodiya, “Wearable internet of things: Concept, architectural components and promises for person-centered healthcare,” In: 4th International Conference on Wireless Mobile Communication and Healthcare-Transforming Healthcare Through Innovations in Mobile and Wireless Technologies (Mobihealth), IEEE, 2014, pp. 304-307.

[62] H. Y. Yang and Z. F. Cui, “Unique journal: Bio-design and manufacturing,” Bio-Design and Manufacturing, vol. 1(1), pp. 1–1, 2018.

[63] M. Resman, M. Pipan, M. Šimic, and N. Herakovič, “A new architecture model for smart manufacturing: A performance analysis and comparison with the RAMI 4.0 reference model,” Advances in Production Engineering and Management, vol.14(2), pp. 153-165, 2019.

[64] P. Singh, and K. Rajbir, "Implementation of the Quality-of-Service Framework Using IoT-Fog Computing for Smart Healthcare." In Real-Life Applications of the Internet of Things, Apple Academic Press, 2022, pp. 397-419.

[65] S. Patel, H. Park, P. Bonato, L. Chan, and M. Rodgers, “A review of wearable sensors and systems with application in rehabilitation,” Journal of neuro engineering and rehabilitation, vol. 9(1), pp. 1-17, 2012.

[66] A. Holzinger, C. Rocker, and M. Ziefle. ¨ From Smart Health to Smart Hospitals,” In Smart health, Springer International Publishing, Cham, 2015, pp. 1-20.

[67] M. M Alam, H. Malik, M. I. Khan, T. Pardy, A. Kuusik, and Y. Le Moullec, “A survey on the roles of communication technologies in IoT-based personalized healthcare applications,” IEEE Access, vol. 6, pp. 36611-36631, 2018.

[68] J. P. Lafleur, A. Jönsson, S. Senkbeil, and J. P. Kutter, ‘‘Recent advances in lab-on-a-chip for biosensing applications,’’ Biosensors and Bioelectronics, vol. 76, pp. 213–233, 2016.

[69] M. A. Serhani, A. Benharref, and E. Badidi, “Towards dynamic non-obtrusive health monitoring based on SOA and cloud,” In: International Conference on Health Information Science, Springer, 2013, pp. 125-136.

[70] W. M. Sweileh, S. W. Al-Jabi, A. S. AbuTaha, H. Z. Sa’ed, F. M. Anayah, and A. F. Sawalha, “Bibliometric analysis of worldwide scientific literature in mobile-health: 2006–2016,” BMC Medical Informatics and Decision Making, vol. 17(1), pp. 1-12, 2017.

[71] Y. Zhang, M. Qiu, C. W. Tsai, M. M. Hassan, and A. Alamri, “Health-CPS: Healthcare cyber-physical system assisted by cloud and big data,” IEEE Systems Journal, vol. 11(1), pp. 88–95, 2015.

[72] K. Guk, G Han, J. Lim, K. Jeong, T. Kang, E. K. Lim, and J. Jung, "Evolution of wearable devices with real-time disease monitoring for personalized healthcare," Nanomaterials, vol. 9(6), pp. 1-23, 2019.

[73] A. Corradi, L. Foschini, C. Giannelli, R. Lazzarini, C. Stefanelli, M. Tortonesi, and G. Virgilli, “Smart appliances and RAMI 4.0: management and servitization of ice cream machines,” IEEE Transactions on Industrial Informatics, vol. 15(2), pp.1007-1016, 2018.

[74] M. Hankel, and B. Rexroth, "Industrie 4.0: The Reference Architectural Model Industrie 4.0 (RAMI 4.0),” ZVEI, vol 2(2), pp. 4-9, 2015.

[75] F. Fraile, R. Sanchis, R. Poler, and A. Ortiz, “Reference models for digital manufacturing platforms,” Applied Sciences, vol. 9(20), pp. 1-25, 2019.

[76] H. Flatt, S. Schriegel, J. Jasperneite, H. Trsek, and H. Adamczyk, “Analysis of the Cyber-Security of industry 4.0 technologies based on RAMI 4.0 and identification of requirements,” In: IEEE 21st International Conference on Emerging Technologies and Factory Automation (ETFA), IEEE, 2016, p. 1-4.

[77] D. Gu, J. Li, X. Li, and C. Liang, “Visualizing the knowledge structure and evolution of big data research in healthcare informatics,” International journal of medical informatics, vol. 98, pp.22-32, 2017.

[78] Platform Industrie 4.0. https://www.plattform-i40.de/I40/Navigation/EN/Home/home.html. Accessed 12 Feb 2019

[79] S. Asghari, and N. J. Navimipour, “Nature inspired meta-heuristic algorithms for solving the service composition problem in the cloud environments,” International Journal of Communication Systems, vol. 31(12), pp. 3708, 2018.

[80] S. Dash, S. K. Shakyawar, M. Sharma, and S. Kaushik, “Big data in healthcare: management, analysis and future prospects,” Journal of Big Data, vol. 6(1), pp. 1-25, 2017.

[81] M. Rath, “Real time analysis based on intelligent applications of big data and IoT in smart health care systems,” International Journal of Big Data and Analytics in Healthcare, vol. 3(2), pp. 45–61, 2018.

[82] I. H. Khan, M. and Javaid, “Big data applications in medical field: A literature review,” Journal of Industrial Integration and Management, vol. 6(01), pp.53-69, 2021.

[83] L. J. Beesley, M. Salvatore, L. G. Fritsche, A. Pandit, A. Rao, C. Brummett, C. J. Willer, L. D. Lisabeth, and B. Mukherjee, “The emerging landscape of health research based on biobanks linked to electronic health records: Existing resources, statistical challenges, and potential opportunities,” Statistics in medicine, vol. 39(6), pp. 773-800, 2020.

[84] B. Lau, S. J. Gange, and R. D. Moore, “Interval and clinical cohort studies: epidemiological issues,” AIDS research and human retroviruses, vol. 23(6), pp. 769–776, 2007.

[85] M. F. Uddin, and N. Gupta, “Seven V's of Big Data understanding Big Data to extract value,” In: Proceedings of the 2014 zone 1 conference of the American Society for Engineering Education, IEEE, 2014, pp. 1-5

[86] D. Ausiello, and S. Shaw, “Quantitative human phenotyping: the next frontier in medicine,” Transactions of the American Clinical and Climatological Association, vol. 125, pp. 219-228, 2014.

[87] C. H. Liu, J. Wen, Q. Yu, B. Yang, and W. Wang, “HealthKiosk: A family-based connected healthcare system for long-term monitoring.” In: IEEE Conference on computer communications workshops (infocom wkshps), 2011, pp. 241-246.

[88] S. Shilo, H. Rossman, and E. Segal, “Axes of a revolution: challenges and promises of big data in healthcare,” Nature medicine, vol. 26(1), pp. 29-38, 2020.

[89] G. Hripcsak, J. D. Duke, N. H. Shah, C. G. Reich, V. Huser, M. J. Schuemie, M. A. Suchard, R. W. Park, I. C. K. Wong, P. R. Rijnbeek, and J. Van Der Lei, “Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers,” Studies in health technology and informatics, vol. 216, pp.574-578, 2015.

[90] U. Raja, D. McManus, J. Hardin, and B. Haynes, “Collaborative rural healthcare network: A conceptual model,” International Journal of Interactive Mobile Technologies, vol. 5(3), pp. 20–23, 2011.

[91] A. K. Triantafyllidis , V. G. Koutkias, I. Chouvarda, and N. Maglaveras, “Development and usability of a personalized sensor-based system for pervasive healthcare. In: 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2014, pp. 6623–6626.

[92] I. S. Kohane, “Using electronic health records to drive discovery in disease genomics,” Nature Reviews Genetics, 12(6), pp. 417–428, 2011.

[93] A. Shankar, A. Dubey, D. Saini, M. Singh, C. P. Prasad, S. Roy, S. J. Bharati, M. Rinki, N. Singh, T. Seth, and M. Khanna, “Environmental and occupational determinants of lung cancer,” Translational lung cancer research, 8(Suppl 1), pp 1-9, 2019

[94] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti, “Detecting novel associations in large data sets.” science, vol. 334(6062), pp.1518-1524, 2011.

[95] P. K. Sahoo, S. K. Mohapatra, and S. L. Wu, "SLA based healthcare big data analysis and computing in cloud network," Journal of Parallel and Distributed Computing, vol. 119, pp. 121–135, 2018.

[96] L. A. Kapsner, M. O. Kampf, S. A. Seuchter, G. Kamdje-Wabo, T. Gradinger, T. Ganslandt, S. Mate, J. Gruendner, D. Kraska, and H. U. Prokosch, "Moving towards an EHR data quality framework: the MIRACUM approach," In German Medical Data Sciences: Shaping Change–Creative Solutions for Innovative Medicine, IOS Press, 2019, pp. 247-253.

[97] S. M, "The quantified self: Fundamental disruption in big data science and biological discovery," Big Data, vol. 1, pp. 85–99, 2013.

[98] T. S. Dahl, M. N. K Boulos, "Robots in health and social care: a complementary technology to home care and telehealthcare?" Robotics, vol. 3 (1), pp. 1–21, 2014.

[99] M. M. Cloutier, P. M. Salo, L. J. Akinbami, R. D. Cohn, J. C. Wilkerson, G. B. Diette, S. Williams, K. S. Elward, J. M. Mazurek, J. R. Spinner, and T. A. Mitchell, "Clinician agreement, self-efficacy, and adherence with the guidelines for the diagnosis and management of asthma," The Journal of Allergy and Clinical Immunology: In Practice, vol. 6(3), pp. 886-894, 2018.

[100] S. Ram, W. Zhang, M. Williams, and Y. Pengetnze, "Predicting Asthma Related Emergency Department visits using big data," IEEE journal of biomedical and health informatics, vol. 19, pp. 1216–1223, 2015.

[101] R. Fang, S. Pouyanfar, Y. Yang, S.-C. Chen, and S. S. Iyengar, "Computational health informatics in the big data age: A survey," ACM Computing Surveys, vol. 49(1) pp. 1-36, 2016.

[102] M. R. Ghazi, and D. Gangodkar, "Hadoop, MapReduce and HDFS: A developers perspective," Procedia Computer Science, vol. 48, pp. 45–50, 2015.

[103] C. Uzunkaya, T. Ensari, Y. Kavurucu, "Hadoop Ecosystem and Its Analysis on Tweets," Procedia - Social and Behavioral Sciences, vol. 195, pp. 1890 – 1897, 2015.

[104] A. O'Driscoll, J. Daugelaite, and R. D. Sleator, "Big data, Hadoop and cloud computing in genomics," Journal of biomedical informatics, vol. 46(5), pp. 774-781, 2013.

[105] H. Huang, T. Gong, N. Ye, R. Wang, and Y. Dou, "Private and secured medical data transmission and analysis for wireless sensing healthcare system," IEEE Transactions on Industrial Informatics, vol. 13(3), pp. 1227-1237, 2017

[106] V. Seethalakshmi, V. Govindasamy, and V. Akila. "Hybrid gradient descent spider monkey optimization (HGDSMO) algorithm for efficient resource scheduling for big data processing in heterogenous environment." Journal of Big Data, vol. 7(1), pp.1-25, 2020.

[107] R. Selvi, Thanga, and I. Muthulakshmi, "Modelling the map reduce based optimal gradient boosted tree classification algorithm for diabetes mellitus diagnosis system," Journal of Ambient Intelligence and Humanized Computing, 12(2), pp.1717-1730, 2021.

[108] N. Yuvaraj, and K. R. SriPreethaa, "Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster," Cluster Computing, 22(1), pp. 1-9, 2019.

[109] R. Ramani, K. Vimala Devi, and K. Ruba Soundar, "MapReduce-based big data framework using modified artificial neural network classifier for diabetic chronic disease prediction," Soft Computing, vol. 24(21), pp. 16335-16345, 2020.

[110] A. Abdelaziz, M. Elhoseny, A. S. Salama, and A. M. Riad, “A machine learning model for improving healthcare services on cloud computing environment,” Measurement, vol. 119, pp. 117-128, 2018.

[111] M. Elhoseny, K. Shankar, and J. Uthayakumar, “Intelligent diagnostic prediction and classification system for chronic kidney disease,” Scientific reports, vol. 9(1), pp. 1-14, 2019.

[112] H. Polat, H. Danaei Mehr, and A. Cetin, “Diagnosis of chronic kidney disease based on support vector machine by feature selection methods,” Journal of medical systems, vol. 41(4), pp. 1-11, 2017.

[113] K. Harimoorthy, M. Thangavelu, “Multi-disease prediction model using improved SVM-radial bias technique in healthcare monitoring system,” Journal of Ambient Intelligence and Humanized Computing, vol. 12(3), pp. 3715-3723, 2021.

[114] M. Makino, R. Yoshimoto, M. Ono, T. Itoko, T. Katsuki, A. Koseki, M. Kudo, K. Haida, J. Kuroda, R. Yanagiya, and E. Saitoh, “Artificial intelligence predicts the progression of diabetic kidney disease using big data machine learning,” Scientific reports, vol. 9(1), pp. 1-9, 2019.

[115] J. Wang, B. Bao, P. Shen, G. Kong, Y. Yang, X. Sun, G. Ding, B. Gao, C. Yang, M. Zhao, and H. Lin, “Using electronic health record data to establish a chronic kidney disease surveillance system in China: protocol for the China Kidney Disease Network (CK-NET)-Yinzhou Study,” BMJ open, vol. 9(8), pp. 1-8, 2019

[116] P. Arulanthu, and E. Perumal, “An intelligent IoT with cloud centric medical decision support system for chronic kidney disease prediction,” International Journal of Imaging Systems and Technology, vol. 30(3), pp. 815-827, 2020.

[117] P. Asghari, A. M. Rahmani, and S. J. H. Haj, “A medical monitoring scheme and health-medical service composition model in cloud-based IoT platform,” Transactions on Emerging Telecommunications Technologies, vol. 30(6), pp. 1-25, 2019

[118] G. R. Vasquez-Morales, S. M. Martinez-Monterrubio, P. Moreno-Ger, J. A. Recio-Garcia, “Explainable prediction of chronic renal disease in the colombian population using neural networks and case-based reasoning,” IEEE Access, vol. 7, pp. 152900-152910, 2019.

[119] H. Kriplani, B. Patel, S. Roy, “Prediction of chronic kidney diseases using deep artificial neural network technique,” In: Computer aided intervention and diagnostics in clinical and medical images, Springer International Publishing, Cham, 2019, pp 179-187.

[120] O. A. Jongbo, A. O. Adetunmbi, R. B. Ogunrinde, B. Badeji-Ajisafe, “Development of an ensemble approach to chronic kidney disease diagnosis,” Scientific African, vol. 8, pp. 1-8, 2020.

[121] B. Boukenze, A. Haqiq, H. Mousannif, “Predicting chronic kidney failure disease using data mining techniques,” In International Symposium on Ubiquitous Networking, Springer, 2016, pp. 701-712.

[122] S. Sossi Alaoui, B. Aksasse, and Y. Farhaoui, “Statistical and predictive analytics of chronic kidney disease,” In: International conference on advanced intelligent systems for sustainable development, Springer, 2018, pp. 27-38.

[123] J. Al-Jaroodi, N. Mohamed, and E. Abukhousa, “Health 4.0: on the way to realizing the healthcare of the future,” IEEE Access, vol. 8, pp. 211189-211210, 2020.

[124] A. B. Lumsden, and J. Bismuth, “Current status of endovascular catheter robotics,” The Journal of Cardiovascular Surgery, vol. 59(3), pp. 310-316, 2018.

[125] H. Melkas, L. Hennala, S. Pekkarinen, and V. Kyrki, “Impacts of robot implementation on care personnel and clients in elderly-care institutions,” International Journal of Medical Informatics, vol.134, pp. 1-6, 2020.

[126] J. Salmeron-Garcia, P. Inigo-Blasco, F. Diaz-del-Rio, and D. Cagigas-Muniz, “Mobile robot motion planning based on cloud computing stereo vision processing,” In: 41st International Symposium on Robotics, VDE, 2014, pp. 1-6.

[127] H. Marcus, D. Nandi, A. Darzi, and G. Z. Yang, “Surgical robotics through a keyhole: From today’s translational barriers to tomorrow’s “disappearing” robots,” IEEE Transaction Biomedical Engineering, vol. 60(3), pp. 674–681, 2013.

[128] G. A. Antoniou, C. V. Riga, E. K. Mayer, N. J. Cheshire, and C. D. Bicknell, “Clinical applications of robotic technology in vascular and endovascular surgery,” Journal of vascular surgery, vol. 53(2), pp. 493-499, 2011.

[129] J. A. Hawks, J. Kunowski, and S. R. Platt, “In vivo demonstration of surgical task assistance using miniature robots,” IEEE Transaction Biomedical Engineering, vol. 59(10), pp. 2866–2873, 2012.

[130] J. Whitman, M. P. Fronheiser, and S.W. Smith, “3-D ultrasound guidance of surgical robotics using catheter transducers: Feasibility study,” IEEE Transaction Ultrasonics, Ferroelectrics, Frequency Control, vol. 55(5), pp. 1143–11435, 2008.

[131] L. J. de Vries, F. Zijlstra, and T. Szili-Torok, “Beyond catheter tip and radiofrequency lesion delivery: The role of robotics in ablation of ventricular tachycardia,” Netherlands Heart journal, vol. 23(10), pp. 483–484, 2015.

[132] H. R. Tari, C. J. Payne, and G. Z. Yang, “Current and Emerging Robot-Assisted Endovascular Catheterization Technologies: A Review,” Annals Bio medical Engineering, vol. 42, pp. 697–715, 2014.

[133] C. Patrone, A. Cella, C. Martini, S. Pericu, R. Femia, A. Barla, C. Porfirione, M. Puntoni, N. Veronese, F. Odone, and N. Casiddu, “Development of a smart post-hospitalization

facility for older people by using domotics, robotics, and automated tele-monitoring," Geriatric Care, vol. 5(1), pp. 12-17, 2019.

[134] W. Deng, I. Papavasileiou, Z. Qiao, W. Zhang, K. Y. Lam, and S. Han, "Advances in automation technologies for lower extremity neurorehabilitation: A review and future challenges," IEEE reviews in biomedical engineering, vol. 11, pp.289-305, 2018.

[135] M. McKnight, "IOT, Industry 4.0, Industrial IOT… Why connected devices are the future of design," KnE Engineering, vol. 2(2), pp.197-202, 2017.

[136] S. Movassaghi, M. Abolhasan, J. Lipman, D. Smith, and A. Jamalipour, "Wireless body area networks: A survey," IEEE Communications Surveys & Tutorials, vol. 16 (3), pp. 1658-1686, 2014.

[137] C. O. Rolim, F. L. Koch, C. B. Westphall, J. Werner, A. Fracalossi, and G. S. Salvador, "A cloud computing solution for patient's data collection in health care institutions," In" Second International Conference on eHealth, Telemedicine, and Social Medicine, IEEE, 2010, pp. 95-99.

[138] Z. Wang, "Detection and Automation Technologies for the Mass Production of Droplet Biomicro fluidics." IEEE Reviews in Biomedical Engineering, vol. 11, pp, 260-274, 2018.

[139] F. Almada-Lobo, "The industry 4.0 revolution and the future of manufacturing execution systems (mes)," Journal of Innovation Management, vol. 3(4), pp. 16-21, 2016.

[140] B. Kehoe, S. Patil, P. Abbeel, and K. Goldberg, "A survey of research on cloud robotics and automation," IEEE Transactions on Automation Science and Engineering, vol. 12(2), pp. 398-409, 2015.

[141] A. K. Gupta, and K. S. Mann, "Sharing of medical information on cloud platform-a review," IOSR Journal of Computer Engineering, vol. 16(2), pp, 08–11, 2014.

[142] D. Ravi, C. Wong, B. Lo, and G. Z. Yang, "A deep learning approach to on-node sensor data analytics for mobile or wearable devices," IEEE Journal of Biomedical and Health Informatics, vol. 21(1), pp. 56–64, 2017.

[143] O. Ali, A. Shrestha, J. Soar, and S. F. Wamba, "Cloud computing-enabled healthcare opportunities, issues, and applications: A systematic review," International Journal of Information Management, vol. 43, pp. 146–158, 2018.

[144] D. Chen, Y. Chen, B. N. Brownlow, P. P. Kanjamala, C. A. G. Arredondo, B. L. Radspinner, and M. A. Raveling, "Real-time or near real-time persisting daily healthcare data into HDFS and elastic search index inside a big data platform," IEEE Transactions on Industrial Informatics, vol. 13(2), pp. 595-606, 2017.

[145] M. Li, S. Yu, Y. Zheng, K. Ren, and W. Lou, "Scalable and secure sharing of personal health records in cloud computing using attribute-based encryption," IEEE transactions on parallel and distributed systems, vol. 24(1), pp. 131-143, 2013.

[146] N. Sultan, “Making use of cloud computing for healthcare provision: Opportunities and challenges,” International Journal of Information Management, vol. 34(2), pp. 177-184, 2014.

[147] C. Doukas, T. Pliakas, and I. Maglogiannis, “Mobile healthcare information management utilizing cloud computing and android OS,” In: Annual International Conference of the IEEE Engineering in Medicine and Biology, 2010, pp. 1037-1040

[148] C. He, X. Jin, Z. Zhao, and T. Xiang, “A cloud computing solution for hospital information system,” In: IEEE International Conference on Intelligent Computing and Intelligent Systems (ICIS), 2010, pp. 517-520.

[149] Z. R. Li, E. C. Chang, K. H. Huang, and F. Lai, “A secure electronic medical record sharing mechanism in the cloud computing platform,” In: IEEE 15th international symposium on consumer electronics (ISCE), 2011, pp. 98-103

[150] C. T. Yang, L. T. Chen, W. L. Chou, and K. C. Wang, “Implementation of a medical image file accessing system on cloud computing,” In: IEEE 13th International Conference on Computational Science and Engineering (CSE), 2010, 321-326.

[151] B. M. Silva, J. J. Rodrigues, I. de la Torre Díez, M. López-Coronado and K. Saleem, “Mobile-health: A review of current state in 2015,” Journal of biomedical informatics, vol. 56, pp.265-272, 2015.

[152] A. Awad, A. Mohamed, C. F. Chiasserini, T. Elfouly, “Distributed in-network processing and resource optimization over mobile-health systems,” Journal of Network Computer Application, vol. 82, pp. 65–76. 2017.

[153] R. S. Istepanaian, and Y. T. Zhang, “Guest editorial introduction to the special section: 4G health—the long-term evolution of m-health,” IEEE Transactions on information technology in biomedicine, vol. 16(1), pp.1-5, 2012.

[154] D. M. West, “How 5G technology enables the health internet of things,” Brookings Center for Technology Innovation, vol. 3(1), pp. 1-20, 2016.

[155] A. Al-Marridi, A. Mohamed, and A. Erbad, “Ai-based techniques on edge devices to optimize energy efficiency in m-health applications,” In Energy Efficiency of Medical Devices and Healthcare Applications, Academic Press, 2020, pp. 1-23.

[156] U. Varshney, “Pervasive healthcare and wireless health monitoring,” Mobile Networks and Applications, vol. 12 (2), pp. 113–127, 2007.

[157] C. Tan, L. Sun, and K. Liu, “Big data architecture for pervasive healthcare: a literature review,” In: Proceedings of the Twenty-third European Conference on Information Systems (ECIS), 2015, pp. 117.

[158] Á. A. D. C. C. Sobrinho, L. D. da Silva, L. M. de Medeiros, and A. C. de Brito Câmara, “yy899,” In: Handbook of Research on ICTs and Management Systems for Improving Efficiency in Healthcare and Social Care, IGI Global, pp. 416-430.

[159] S. P Ruotsalainen, B. Blobel, A. Seppala, P. Nykanen, “Trust information-based privacy architecture for ubiquitous health,” JMIR Mhealth and Uhealth, vol. 1(2), pp. 1-15 ,2013

[160] F. Touati, and R. Tabish, “U-healthcare system: State-of-the-art review and challenges,” Journal of medical systems, vol. 37(3), pp.1-20, 2013.

[161] H. Chang, M. Choi, “Big data and healthcare: building an augmented world,” Health. Informatics Research, vol. 22 (3), pp.153–155, 2016.

[162] E. Omanovi´c-Mikliˇcanin, M. Maksimovi´c, V. Vujovi´c, “The future of healthcare: nanomedicine and internet of nano things,” Folia Medica Facultatis Medicinae Universitatis Saraeviensis, vol. 50 (1), pp.1-3, 2015.

[163] Benharref, Abdelghani, and Mohamed Adel Serhani, "Novel cloud and SOA-based framework for E-Health monitoring using wireless biosensors," IEEE journal of biomedical and health informatics, vol.18(1), pp. 46-55, 2013.

[164] A. Kumari, S. Tanwar, S. Tyagi, N. and Kumar, “Fog computing for Healthcare 4.0 environment: Opportunities and challenges,” Computers & Electrical Engineering, vol. 72, pp.1-13, 2018.

[165] G. Manogaran, C. Thota, D. Lopez, and R. Sundarasekar, “Big data security intelligence for healthcare industry 4.0”. In Cybersecurity for industry 4.0, Springer, 2017, pp. 103-126.

[166] D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar, “Big data in health care: using analytics to identify and manage high-risk and high-cost patients,” Health affairs, vol. 33(7), pp. 1123-1131, 2014.

[167] C. Thuemmler, and C. Bai, “Health 4.0: application of industry 4.0 design principles in future asthma management,” In Health 4.0: How virtualization and big data are revolutionizing healthcare, Springer, 2017, pp. 23-37

[168] S. González-Valenzuela, M. Chen, and V. C. Leung, “Mobility support for health monitoring at home using wearable sensors,” IEEE Transactions on Information Technology in Biomedicine, vol. 15(4), pp. 539-549, 2011.

[169] M. Viceconti, P. Hunter, and R. Hose, “Big data, big knowledge: big data for personalized healthcare,” IEEE journal of biomedical and health informatics, vol. 19(4), pp. 1209-1215, 2015.

[170] G. Manogaran, and D. Lopez, “Health data analytics using scalable logistic regression with stochastic gradient descent,” International Journal of Advanced Intelligence Paradigms, 10(1-2), pp.118-132, 2018.

[171] N. V. Chawla, and D. A. Davis, “Bringing big data to personalized healthcare: a patient-centered framework,” Journal of general internal medicine, vol. 28(3), pp.660-665, 2013.

[172] H. Dubey, J. C. Goldberg, M. Abtahi, L. Mahler, and K. Mankodiya, “EchoWear: smartwatch technology for voice and speech treatments of patients with Parkinson's disease,” In: Proceedings of the conference on Wireless Health, ACM, 2015, pp. 1-8.

[173] O. Akrivopoulos, I. Chatzigiannakis, C. Tselios, and A. Antoniou, “On the deployment of healthcare applications over fog computing infrastructure,” In: IEEE 41st annual computer software and applications conference (COMPSAC), IEEE, 2017, pp. 288-293.

[174] Laine, Teemu H., Chaewoo Lee, and Haejung Suk, "Mobile gateway for ubiquitous health care system using zigbee and bluetooth." In 2014 Eighth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, pp. 139-145. IEEE, 2014.

[175] A. Bourouis, A. Zerdazi, M. Feham, and A. Bouchachia, “M-health: skin disease analysis system using smartphone's camera,” Procedia Computer Science, vol. 19, pp.1116-1120, 2013.

[176] P. Kumar, and K. Silambarasan, “Enhancing the performance of healthcare service in IoT and cloud using optimized techniques,” IETE Journal of Research, vol. 68(2), pp.1475-1484, 2019.

[177] A. M. Elmisery, S. Rho, and D. Botvich, “A fog based middleware for automated compliance with OECD privacy principles in internet of healthcare things,” IEEE Access, vol. 4, pp.8418-8441, 2016.

[178] H. Dubey, J. Yang, N. Constant, A. M. Amiri, Q. Yang, and K. Makodiya, “Fog data: Enhancing telehealth big data through fog computing,” In: Proceedings of the ASE bigdata & socialinformatics, ACM, 2015, pp. 1-6.

[179] C. A. Da Costa, C. F. Pasluosta, B. Eskofier, D. B. Da Silva, and R. da Rosa Righi, “Internet of health things: toward intelligent vital signs monitoring in hospital wards,” Artificial intelligence in medicine, vol. 89, pp.61-69, 2018.

[180] V. V. Estrela, A. C. B. Monteiro, R. P. França, Y. Iano, A. Khelassi, and N. Razmjooy, “Health 4.0: applications, management, technologies and review,” Medical Technologies Journal, vol. 2(4), pp. 262-276, 2018.

[181] I. Chiuchisan, I. Chiuchisan, and M. Dimian, “Internet of Things for e-Health: An approach to medical applications,” In: International Workshop on Computational Intelligence for Multimedia Understanding (IWCIM), IEEE, 2015, pp. 1-5.

[182] M. Arlotti, M. Rosa, S. Marceglia, S. Barbieri, and A. Priori, “The adaptive deep brain stimulation challenge,” Parkinsonism & related disorders, vol. 28, pp.12-17, 2016.

[183] H. Löhr, A. R. Sadeghi, and M. Winandy, “Securing the e-health cloud,” In: Proceedings of the 1st acm international health informatics symposium, 2010, pp. 220-229.

[184] A. M. Rahmani, T. N. Gia, B. Negash, A. Anzanpour, I. Azimi, M. Jiang, and P. Liljeberg, “Exploiting smart e-Health gateways at the edge of healthcare Internet-of-Things: A fog computing approach,” Future Generation Computer Systems, vol. 78, pp. 641-658, 2018.

[185] G. Suciu, V. Suciu, A. Martian, R. Craciunescu, A. Vulpe, I. Marcu, S. Halunga, and O. Fratu, “Big data, internet of things and cloud convergence–an architecture for secure e-health applications,” Journal of medical systems, vol. 39(11), pp.1-8, 2015.

[186] F. Firouzi, B. Farahani, M. Ibrahim, and K. Chakrabarty, “Keynote paper: from EDA to IoT eHealth: promises, challenges, and solutions,” IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 37(12), pp.2965-2978, 2018.

[187] M. A. d. Jesus, and V. V. Estrela. "An introduction to data mining applied to health-oriented databases." Oriental Journal of Computer Science and Technology, vol 9(3), 2016.

[188] A. S. Goy, V. D. Nishtar, C. Balatbat, and R. Diabo. "Health and healthcare in the Fourth Industrial Revolution: Global Future Council on the Future of Health and Healthcare 2016-2018." World Economic Forum, 2019. Accessed on 2. March.2019.

[189] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," IEEE internet of things journal, vol.3(5), pp. 637-646, 2016.

[190] R. Gravina, M. C. P. Pace, G. Aloi, W. Russo, W. Li, and G. Fortino, "Cloud-based Activity-aaService cyber–physical framework for human activity monitoring in mobility," Future Generation Computer Systems, vol. 75, pp. 158-171, 2017.

[191] G. Sannino, and G. D. Pietro, G. "A deep learning approach for ECG-based heartbeat classification for arrhythmia detection," Future Generation Computer Systems, vol. 86, pp. 446-455, 2018.

[192] S. Sengupta and S. S. Bhunia, "Secure Data Management in Cloudlet Assisted IoT Enabled e-Health Framework in Smart City," IEEE Sensors Journal, vol. 20(16), pp. 9581-9588, 2020.

[193] C. Yvanoff-Frenchin, V. Ramos, T. Belabed, and C. Valderrama, "An Edge Computing Robot Experience for Automatic Elderly Mental Health Care Based on Voice," arXiv, pp. 1-9, 2019.

[194] World Health Organization, "Global status report on water safety plans: a review of proactive risk assessment and risk management practices to ensure the safety of drinking-water, 2017, (No. WHO/FWC/WSH/17.03).

[195] M. Ghassemi, L. A. Celi, and D. J. Stone, "State of the art review: the data revolution in critical care," Critical Care, vol. 19(1), pp.1-9, 2015.

[196] N. Agarwal, and A. Brem, "Strategic business transformation through technology convergence: implications from General Electric's industrial internet initiative," International Journal of Technology Management, vol. 67(2-4), pp. 196-214, 2015.

[197] M. Kumar, J. Mostafa, and R. Ramaswamy, "Federated health information architecture: enabling healthcare providers and policymakers to use data for decision-making," Health Information Management Journal, vol. 47(2): pp. 85-93, 2018.

[198] M. King, A. Smith, and M. Gracey, "Indigenous health part 2: the underlying causes of the health gap," The lancet, vol. 374(9683), pp. 76-85, 2009.

[199] A. MOHYUDDIN, and I. REHMAN, "Economic issues of senior citizens," European Academic Research, vol. 3(6), pp. 7050-7064, 2015.

[200] C. Lahariya, "Ayushman Bharat'program and universal health coverage in India," Indian pediatrics, vol. 55(6), pp. 495-506, 2018.

[201] World Health Organization, Diet, nutrition, and the prevention of chronic diseases: report of a joint, WHO/FAO expert consultation, Vol. 916, 2003, World Health Organization

[202] T. B. Murdoch, and A. S. Detsky, "The inevitable application of big data to health care," Jama, vol. 309(13), pp. 1351-1352, 2013.

[203] T. Alsuliman, D. Humaidan, and L. Sliman, "Machine learning and artificial intelligence in the service of medicine: Necessity or potentiality?," Current research in translational medicine, vol. 68(4), pp. 245-251, 2020.

[204] L. E. Bayne, "Big data in neonatal health care: big reach, big reward?," Critical Care Nursing Clinics, vol. 30(4), pp. 481-497, 2018.

[205] A. K. Waljee, R. Lipson, W. L. Wiitala, Y. Zhang, B. Liu, J. Zhu, B. Wallace, S. M. Govani, R. W. Stidham, R. Hayward, and P. D. Higgins, "Predicting hospitalization and outpatient corticosteroid use in inflammatory bowel disease patients using machine learning,". Inflammatory bowel diseases, vol. 24(1), pp. 45-53, 2018.

[206] A. Belle, R. Thiagarajan, S. M. Soroushmehr, F. Navidi, D. A. Beard, and K. Najarian, "Big data analytics in healthcare," BioMed research international, vol. 2015, pp. 1-17, 2015.

[207] A. Jacobs, "The pathologies of big data," Communications of the ACM, Vol. 52(8), pp. 36–44, 2009.

[208] N. Mehta, and A. Pandit, "Concurrence of big data analytics and healthcare: A systematic review," International journal of medical informatics, vol. 114, pp.57-65, 2018.

[209] P. Galetsi, and K. Katsaliaki, "A review of the literature on big data analytics in healthcare," Journal of the Operational Research Society, vol. 71(10), pp.1511-1529, 2020.

[210] E. Mbunge, B. Muchemwa, and J. Batani, "Sensors and healthcare 5.0: transformative shift in virtual care through emerging digital health technologies," Global Health Journal, vol. 5(4), pp. 169-177, 2021.

[211] M. Elhoseny, A. Abdelaziz, A. S. Salama, A. M. Riad, K. Muhammad, and A. K. Sangaiah, "A hybrid model of internet of things and cloud computing to manage big data in health services applications," Future generation computer systems, vol. 86, pp. 1383-1394, 2018.

[212] F. D. Silveira, I. R. Neto, F. M. Machado, M.P. D. Silva, and F. G. Amaral, "Analysis of industry 4.0 technologies applied to the health sector: systematic literature review," Occupational and environmental safety and health, Springer, 2019, pp.701-709.

[213] H. Nasiri, S. Nasehi, and M. Goudarzi, "Evaluation of distributed stream processing frameworks for IoT applications in Smart Cities," Journal of Big Data, vol. 6(1), pp.1-24, 2019.

[214] D. Sonntag, V. Tresp, S. Zillner, A. Cavallaro, M. Hammon, A. Reis, P. A. Fasching, M. Sedlmayr, T. Ganslandt, H. U. Prokosch, K. Budde, D. Schmidt, C. Hinrichs, T. Wittenberg, P. Daumke and P. G. Oppelt, "The Clinical Data Intelligence Project: A Smart Data Initiative," Informatik Spektrum, vol. 39, pp.290- 300, 2016.

[215] K. Budde, T. Becker, W. Arns, C. Sommerer, P. Reinke, U. Eisenberger, S. Kramer, W. Fischer, H. Gschaidmeier and F Pietruck, "Everolimus-based, calcineurin-inhibitor free regimen in recipients of de-novo kidney transplants: an open-label, randomised, controlled trial," Lancet, vol. 377, pp.837–847, 2011.

[216] K. Budde, F. Lehner, C. Sommerer, W. Arns, P. Reinke, U. Eisenberger, R. P. Wüthrich, S. Scheidl, C. May, E. M. M. Paulus, A. Mühlfeld, H. H. Wolters, K. Pressmar, R. Stahl and O. Witzke, "Conversion from cyclosporine to everolimus at 4.5 months post-transplant: 3-year results from the randomized ZEUS study" American J. Transplant, vol. 12, pp.1528–1540, 2012

[217] C. Hinrichs, S. Wendland, H. Zimmermann, D. Eurich, R. Neuhaus, P. Schlattmann, N. Babel, H. Riess, B. Gärtner, I. Anagnostopoulos, P. Reinke, R. U. Trappe, "IL-6 and IL-10 in post-transplant lymphoproliferative disorders development and maintenance: a longitudinal study of cytokine plasma levels and T-cell subsets in 38 patients undergoing treatment," Transplant International, Vol. 24, pp. 892-903, 2011.

[218] L. Huber, M, Naik and K. Budde, "Desensitization of HLA Incompatible Kidney Recipients," New Engl. J. Med., vol. 365, pp. 643–1645, 2011.

[219] J. Lasserre, S. Arnold, M. Vingron, P. Reinke and C. Hinrichs, "Predicting the outcome of renal transplantation," JAMIA, vol. 19, pp. 255–262, 2012.

[220] N. Cozzoli, F. P. Salvatore, N. Faccilongo, and M. Milone, "How can big data analytics be used for healthcare organization management? Literary framework and future research from a systematic review," BMC health services research, vol. 22(1), pp. 1-14, 2022.

[221] https://data.gov.in

[222] V. Tresp, J. M. Overhage, M. Bundschus, S. Rabizadeh, P. A. Fasching, and S. Yu, "Going digital: a survey on digitalization and large-scale data analytics in healthcare," Proceedings of the IEEE, vol. 104(11), pp 2180-2206, 2016.

[223] V. Kumar, D. R. Recupero, D. Riboni, and R. Helaoui, "Ensembling classical machine learning and deep learning approaches for morbidity identification from clinical notes," IEEE Access, vol. 9, pp. 7107-7126, 2020.

[224] A. Nishanth, and T. Thiruvaran, "Identifying important attributes for early detection of chronic kidney disease," IEEE reviews in biomedical engineering, vol. 11, pp. 208-216, 2017.

[225] V. Kumar, B. K. Mishra, M. Mazzara, D. N. Thanh, and A. Verma, "Prediction of malignant and benign breast cancer: A data mining approach in healthcare applications," In: Advances in data science and management, Springer, 2020, pp. 435-442.

[226] Z. Zhang, B. Wang, F. Ahmed, I. V. Ramakrishnan, R. Zhao, A. Viccellio, and K. Mueller, "The five Ws for information visualization with application to healthcare informatics," IEEE transactions on visualization and computer graphics, vol. 19(11), pp. 1895-1910, 2013.

[227] https://www.niemanlab.org/2015/12/the-five-es-of-journalism-in-2016/

[228] T. White, “Hadoop: The definitive guide,” " O'Reilly Media, Inc.", 2012, pp. 45-194.

[229] E. Hwang, S. Kim, J. S. Kim, S. Hwang, and Y. R. Choi, “On the role of application and resource characterizations in heterogeneous distributed computing systems,” Cluster Computing, vol. 19(4), pp. 2225-2240, 2016.

[230] G. Manogaran, and D. Lopez, “Disease surveillance system for big climate data processing and dengue transmission,” In: Climate Change and Environmental Concerns: Breakthroughs in Research and Practice, IGI Global, 2018, pp. 427-446.

[231] C. W. Lee, K. Y. Hsieh, S. Y. Hsieh, and H. C. Hsiao, “A dynamic data placement strategy for Hadoop in heterogeneous environments,” Big Data Research, vol. 1, pp. 14-22, 2014.

[232] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, “Using collaborative filtering to weave an information tapestry,” Communications of the ACM, vol. 35(12), pp. 61-70, 1992.

[233] P. Resnick, and H. R. Varian, “Recommender systems,” Communications of the ACM, vol. 40(3), pp. 56-58, 1997.

[234] J. S. Breese, D. Heckerman, and C. Kadie, “Empirical analysis of predictive algorithms for collaborative filtering,” In: Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence, 2013, pp. 43-52.

[235] D. Dua, and C. Graff, “UCI machine learning repository,” Irvine, CA: University of California, School of Information and Computer Science, 2017.

[236] A. V. Tikoo, “Transplantation of human organs: The Indian scenario,” ILI law review, vol. 1, pp. 147-174, 2017.

[237] Y. Ren, H. Fei, X. Liang, D. Ji, and M. Cheng, “A hybrid neural network model for predicting kidney disease in hypertension patients based on electronic health records,” BMC medical informatics and decision making, vol. 19(2), pp. 131-138, 2019.

[238] O. R. G. A. N. India, “A study of the deceased organ donation environment in Delhi/NCR,” An initiative of the Parashar Foundation in partnership with MOHAN Foundation. Outline India, 2014, pp. 1-61

[239] N. Jose, “Challenges in Organ Transplantation-An Indian Scenario,” Indian Journal of Surgical Nursing, vol. 5(1), pp. 9, 2016.

[240] R. J. P. Princy, S. Parthasarathy, P. S. H. Jose, A. R. Lakshminarayanan, and S. Jeganathan, “Prediction of cardiac disease using supervised machine learning algorithms,” In: 4th International Conference on Intelligent Computing and Control Systems (ICICCS), IEEE, 2020, pp. 570-575.

[241] C. X. Ling, J. Huang, and H. Zhang, “AUC: a better measure than accuracy in comparing learning algorithms,” In: Conference of the canadian society for computational studies of intelligence, Springer, 2003, pp. 329-341.

[242] Ling, C. X., Huang, J., & Zhang, H. (2003, August). AUC: a statistically consistent and more discriminating measure than accuracy. In Ijcai, 3, pp. 519-524.

[243] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," Pattern recognition, vol. 30(7), pp. 1145-1159, 1997.

[244] X. Li, J. Zhang, and F. Safara, "Improving the accuracy of diabetes diagnosis applications through a hybrid feature selection algorithm," Neural Processing Letters, pp.1-17, 2021.

[245] M. Langarizadeh, and F. Moghbeli, "Applying naive bayesian networks to disease prediction: a systematic review," Acta Informatica Medica, vol. 24(5), pp. 364, 2016.

[246] J. Yu, P. Ping, L. Wang, L. Kuang, X. Li, and Z. Wu, "A novel probability model for lncRNA–disease association prediction based on the naïve bayesian classifier," Genes, vol. 9(7), pp. 345, 2018.

[247] A. Darwish, A. E. Hassanien, M. Elhoseny, A. K. Sangaiah, and K. Muhammad, "The impact of the hybrid platform of internet of things and cloud computing on healthcare systems: opportunities, challenges, and open problems," Journal of Ambient Intelligence and Humanized Computing, vol. 10(10), pp. 4151-4166, 2019.

[248] G. Aceto, V. Persico, A. Pescapé, "Industry 4.0 and health: Internet of things, big data, and cloud computing for healthcare 4.0.," Journal of Industrial Information Integration, vol. 1(18), pp. 1-14, 2020.

[249] L. Thames, and D. Schaefer, "Cybersecurity for Industry 4.0 and advanced manufacturing environments with ensemble intelligence," In Cybersecurity for Industry 4.0 , Springer, 2017, pp. 243-265.

[250] F. Zezulka, P. Marcon, I. Vesely, and O. Sajdl, "Industry 4.0–An Introduction in the phenomenon," IFAC-PapersOnLine, vol. 49(25), pp. 8-12, 2016.

[251] G. Manogaran, and D. Lopez, "A survey of big data architectures and machine learning algorithms in healthcare," International Journal of Biomedical Engineering and Technology, vol. 25(2-4), pp.182-211, 2017

[252] P. M. Dhulavvagol, S. G. Totad, and S. Sourabh, "Performance analysis of job scheduling algorithms on Hadoop multi-cluster environment," In Emerging Research in Electronics, Computer Science and Technology, Springer, 2019, pp. 457-470.

[253] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, and J. Seekins, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," In: Proceedings of the AAAI conference on artificial intelligence, 2019, Vol. 33(01), pp. 590-597.

[254] S. Wang, J. Wan, D. Zhang, D. Li, and C. Zhang, "Towards smart factory for industry 4.0: a self-organized multi-agent system with big databased feedback and coordination," Computer networks, vol. 101, pp. 158-168, 2016.

[255] C. Salkin, M. Oner, A. Ustundag, and E. Cevikcan, "A conceptual framework for Industry 4.0.," In Industry 4.0: managing the digital transformation, Springer, 2018, pp. 3-23.

[256] L. Wang, and G. Wang, “Big data in cyber-physical systems, digital manufacturing and industry 4.0.,” International Journal of Engineering and Manufacturing (IJEM), vol. 6(4), pp.1-8, 2016.

[257] C. Wittenberg, “Cause the trend Industry 4.0 in the automated industry to new requirements on user interfaces?,” In: International Conference on Human-Computer Interaction, Springer, 2015, pp. 238-245.

[258] B. Bagheri, S. Yang, H. A. Kao, and J. Lee, “Cyber-physical systems architecture for self-aware machines in industry 4.0 environment,” IFAC-PapersOnLine, vol. 48(3), pp. 1622-1627, 2015.

[259] M. Javaid, and A. Haleem, “Additive manufacturing applications in medical cases: A literature based review,” Alexandria Journal of Medicine, vol. 54(4), pp. 411-422, 2018.

[260] L. K. Cheung, M. C. Wong, and L. L. Wong, “Refinement of facial reconstructive surgery by stereo-model planning,” Annals of the Royal Australasian College of Dental Surgeons, vol. 16, pp. 129-132, 2002

[261] I. Gibson, L. K. Cheung, S. P. Chow, W. L. Cheung, S. L. Beh, M. Savalani, and S. H. Lee, “The use of rapid prototyping to assist medical applications,” Rapid Prototyping Journal, vol. 12, pp. 53-58, 2006.

[262] J. Tuomi, K. S. Paloheimo, J. Vehviläinen, R. Björkstrand, M. Salmi, E. Huotilainen, R. Kontio, S. Rouse, I. Gibson, and A. A. Mäkitie, “A novel classification and online platform for planning and documentation of medical applications of additive manufacturing,” Surgical innovation, vol. 21(6), pp. 553-559, , 2014.

[263] M. Javaid, and A. Haleem, “Using additive manufacturing applications for design and development of food and agricultural equipments,” International Journal of Materials and Product Technology, vol. 58(2-3), pp. 225-238, 2019.

[264] R. Bibb, D. Eggbeer, P. Evans, A. Bocca, and A. Sugar, “Rapid manufacture of custom-fitting surgical guides,” Rapid Prototyping Journal, vol. 15, pp. 346-354, 2009.

[265] S. R. Sharpton, V. Ajmera, and R. Loomba, “Emerging role of the gut microbiome in nonalcoholic fatty liver disease: from composition to function,” Clinical Gastroenterology and Hepatology, vol. 17(2), pp. 296-306, 2019.

[266] M. Rüßmann, M. Lorenz, P. Gerbert, M. Waldner, J. Justus, P. Engel, and M. Harnisch, “Industry 4.0: The future of productivity and growth in manufacturing industries,” Boston consulting group, vol. 9(1), pp. 54-89, 2015

[267] J. I. Westbrook, J. Braithwaite, R. Iedema, and E. W. Coiera, “Evaluating the impact of information communication technologies on complex organizational systems: a multi-disciplinary, multi-method framework” In: MEDINFO, IOS Press, 2004, pp. 1323-1327.

[268] M. Presser, C. Philippe, T. Christoph, Li. Jun, Z. Xueli, G. Yuming, L.Yang, L. Haihua, G. Xin, and B Chunxue. "EU-China Joint White Paper on the Internet of Things." In: IoT Week Belgrade, 2016, pp. 1-20

[269] K. Vishnevskiy, O. Karasev, and D. Meissner, “Integrated roadmaps for strategic management and planning,” Technological Forecasting and Social Change, vol. 110, pp. 153-166, 2016.

[270] R. Phaal, C. Farrukh, and D. Probert, “Technology Roadmapping: linking technology resources to business objectives,” Centre for Technology Management, University of Cambridge, pp.1-18, 2001.

[271] R. Phaal, C. Farrukh, and D. Probert, “Customizing roadmapping. Research-Technology Management”, vol. 47(2), pp. 26-37, 2004.

[272] R. Phaal, C. J. Farrukh, and D. R. Probert, “Technology roadmapping—A planning framework for evolution and revolution”, Technological forecasting and social change, vol. 71(1-2), pp. 5-26, 2004.

[273] S. Erol, A. Schumacher, and W. Sihn, “Strategic guidance towards Industry 4.0–a three-stage process model,” In: International conference on competitive manufacturing, Vol. 9(1), pp. 495-501, 2016.