An interactive healthcare system of big data application with predictive model analysis

Md. Ataur Rahman Bhuiyan

ID: 2014-1-60-101

Md. Rifat Ullah

ID: 2014-1-60-102

A thesis submitted in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering

Department of Computer Science and Engineering
East West University
Dhaka-1212, Bangladesh

November, 2017

Declaration

We, hereby, declare that the work presented in this thesis is the outcome of the investigation performed by us under the supervision of Amit Kumar Das , Lecturer, Department of Computer Science and engineering, East West University. We also declare that no part of this thesis has been or is being submitted elsewhere for the award of any degree or diploma.

Countersigned

. . . .

(Amit Kumar Das)

Supervisor

Signature

. . . .

(Md. Ataur Rahman Bhuiyan)

(2014-1-60-101)

Signature

. . . .

(Md. Rifat Ullah)

(2014-1-60-102)

Letter of Acceptance

This thesis report entitled *"An interactive healthcare system of big data application with predictive model analysis"* submitted by Md. Ataur Rahman Bhuiyan (ID: 2014-1-60-101) and Md. Rifat Ullah (ID: 2014-1-60-102) to the Department of Computer Science and Engineering, East West University is accepted by the department in partial fulfillment of requirements for the Award of the Degree of Bachelor of Science and Engineering on December, 2017.

Supervisor

. . . .
(Amit Kumar Das)
Lecturer, Department of Computer Science and Engineering, East West University

Chairperson

. . . .
(Dr. Ahmed Wasif Reza)
Chairperson and Associate Professor,
Department of Computer Science and Engineering, East West University

Abstract

Currently, scientific research on healthcare demand to develop an interactive solution to provide healthy life facility with earlier disease detection to the user. In the recent time, healthcare industries are generating lots of unstructured or semi-structured data which needs to be analyzed and processed in real time. In this paper, we have designed a healthcare system to deal with patients biological, and emotional condition as well as the previous health history with genetical data. The data generated by the patient and the hospital are gathered in High-Performance Computer server, and the medical history, as well as genetical data, are collected from the cloud synchronization. We proposed a probabilistic data acquisition scheme to analyze the data and apply MapReduce algorithm in *HPC* to make structure database. The system holds a data warehouse which provides a two-way interaction between *HPC* and cloud for interactive information gathering. In this research, we present a *prediction algorithm* which is performed in cloud server to predict a patients disease. We apply *Random Forest*, *SVM*, *C5.0*, *Naive Bayes* and *Artificial Neural Network* for prediction analysis and shows the side by side comparison on those algorithms.

Acknowledgments

As it is true for everyone, We have also arrived at this point of achieving a goal in our life through various interactions with and help from other people. However, written words are often elusive and harbor diverse interpretations even in one's mother language. Therefore, We would not like to make efforts to find best words to express our thankfulness other than simply listing those people who have contributed to this thesis itself in an essential way. This work was carried out in the Department of Computer Science and Engineering at East West University, Bangladesh.

First of all, We would like to express our deepest gratitude to the almighty Allah for His blessings on us. Next, our special thanks goes to our supervisor, "Amit Kumer Das", who gave us this opportunity, initiated us into the field of "An interactive healthcare system of big data application with predictive model analysis", and without whom this work would not have been possible. His encouragements, visionaries and thoughtful comments and suggestions, unforgettable support at every stage of our B.Sc study were simply appreciating and essential. His ability to muddle us enough to finally answer our own question correctly is something valuable what We have learned and We would try to emulate, if ever We get the opportunity.

We would like to thank "Sadik" for his excellent collaboration during mathematical analysis. For reviewing our paper and correcting grammatical errors, we would like to thank "Hridoy", "Muktasib". We would like to give thanks to "Maruf Islam" for his valuable time with reviewing our paper and giving us suggestions to improve our writing skills during the whole time. We would like to thank “Nadia Kazi” for her encouragement to learn about medical diagnosis system and make the medical term familiar, “Pipasa Amrin Nitu” for her helpful suggestions in understanding medical diagnosis system, and “Nur Uddin Maruf” for his support and help to understand Illustrations works. Last but not the least; we would like to thank our parents for their unending support, encouragement and prayers.

There are numerous other people too who have shown me their constant support and friendship in various ways, directly or indirectly related to our academic life. We will remember them in our heart and hope to find a more appropriate place to acknowledge them in the future.

Md. Ataur Rahman Bhuiyan
November, 2017

Md. Rifat Ullah
November, 2017

Table of Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

1.1 Introduction

As the health consciousness is increased among the people, the medical healthcare industry has become the emerging areas of research and realize the need for a smart and interactive healthcare system which will provide more cost effective and efficient treatment. This is one of the most emerging research areas that attracts many researchers in recent time. Nowadays, medical treatment needs an interactive and intelligent system which can deal with a large biological dataset with human-computer interaction to explore the most valuable information and provide a better treatment related to health. At present, its not incomprehensible to think or design a smart healthcare system which can interact with human very quickly and an effective way. To provide some remarkable facilities related to healthcare, the National Health Reform sets some goals and objectives which monitor the health progress and to identify the necessary changes. The focusing area of broad reach of health services research today is the particular field of the healthcare. This will be expectedly moving towards the advanced health services researchers [1].

1.2 HCI in Data Collection

Biosensor based IoT devices combined with smart wireless technology and data mining technique leads to better healthcare system to explore information with interactive patient monitoring method. This new interactive healthcare system will be able to collect many patients' data (e.g. logical and emotional data of patients) in an interactive way using biosensors based IoT devices and sensor-less devices within a very short period. Researchers consider that, medical data collection is one of the significant portion in a healthcare research [2]. Data collection is enormously necessary to diagnose a patient's disease and exploration of other valuable reports with future disease prediction. Thus, a Human Computer interaction is essential to collect data from patient to ensure greater number of data with less efforts and less time which need to be analyzed.

1.3 Healthcare with Big Data

Medical healthcare systems produce massive volume of continuous medical data and those data come with unstructured or semi-structured format. The existing healthcare systems use electronic health records to store those data. American Hospital Association showed that uses of Electronic Health Records became double from 2009 to 2011 [3]. As reported by the healthcare data analysist, 150 exabytes of medical data are produced by USA healthcare in 2011 [4]. In 2014, this amount is reached to zettabytes [5].

1.3.1 Attributes of Medical Big Data

Interactive healthcare system must be negotiated with this large volume of medical data with big data analytics to discover the hidden patterns and search for unrevealed correlation with the patient's previous medical data. In medical big data analytics, the major challenge is to deal with the attributes of big data which is defined by 5Vs: Volume, Velocity, Variety, Value, and Veracity. In this new interactive healthcare system, patient's data which are collected from various sources describe the volume of data. The arrival rate of data represents the velocity of data. The healthcare data such as ECG, EMG, clinical reports, doctor's note come from many sources and those come with structure, semi-structured and unstructured format that describe the variety. Those data need to be analyzed to find out the meaningful information which is considered as the value of data. There are many uncertain number of state in data which we called hidden data are represent the veracity.

1.3.2 Structure of Big Data

Some advanced technology is used to analyze big data such as tensors, cloud computing and some intelligent framework which can handle missing data from large amount of medical data [6]. But, processing of big data leads to difficulties as there are many semi-structured and unstructured data. IBM healthcare researchers found that, 80% of big data in healthcare are unstructured and they also claim that to retrieve fruitful information about patients, those unstructured or semi-structured data need to be analyzed to make those data in structured format [7]. Another blog named MapR Converged Data Platform notifies that in today's healthcare environment, 75% or more of the data by some estimates is unstructured data [8]. On the other hand, some of the researchers shows that digital data are growing rapidly. They compare the size of the structured and unstructured data of the last decade. According to International Data Corporation (IDC) research, among all the total incremental digital data almost 90% data are unstructured [9]. Sometimes big data known as dirty data and before big data analytics, we must structure those medical data for better analytics result. Unstructured and semi-structured dataset can hold important information about data and those unstructured or semi-structured dataset causes incorrect information about the dataset [10].

1.4 HCI in Big Data Analysis

Big data analytics can deal with a patient's medical historical data, genetic and familial data and emotional data in an effective and in a smarter way along with the help of more efficient operations. It is good for cost-effective and better for good decision making using analyzing result from the data sets. That's the reason big data analytics technology is important for the interactive healthcare system. By analyzing the big amounts of patient's medical information (structured and unstructured), interactive healthcare system can provide lifesaving diagnosis, treatment options with disease prediction of a patient and explore the other functionalities that's lead another Human Computer Interaction for its user.

Another section of interactive healthcare system which needs to be considered is to provide a user-friendly environment in data analysis. HCI in data analytics can activate the data visualization with new technology such as Virtual Reality and Augmented Reality across the data warehouse and give them access to that data over time by users. Data analytics is very indispensable to understand the users and HCI in data analytics can do it effectively and accurately. Advanced HCI solution generates not only confidential data but also represents big data analytics results more effectively in user-friendly environment [11]. HCI in data analytics helps to visualize the graphical view of data to support the analyzing result and it helps the user to know what is going on [3] with the data, which is complex to know only using raw data.

1.5 Prediction Analysis

From all the medical data of a patient using big data analytics and HCI, interactive healthcare system must predict the disease of the patient and should predict probable disease that a patient may suffer in future. There is an excellent relationship between Prediction Model and medical data quality, data volume, and patient outcome data. At the recent year, some researchers experiment about the prediction analysis in medical and describe the importance, also showed that improvement of healthcare by doing prediction analysis [12].

Some researchers claim that prediction model is used for several purposes such as to predict future events of disease and clinical prediction model [13]. Prediction Model helps to minimize treatment variation and unexpected costs. Interactive healthcare system must be standardized with respect to prediction model outcome with other functionalities and the healthcare cost must be reduced. Prediction model within an interactive healthcare system can play a significant role to predict a patient's disease as much as accurately. Using the prediction outcome, doctors don't need to wait for lots of medical test to diagnose a disease. This prediction model helps doctor to communicate with his/her patient very easily. This HCI provide the better service to both doctor and patient by reducing the time of diagnosis and irrelevant cost of treatment.

1.6 Our Contribution

Within this paper, we have implemented an Interactive Healthcare system which can take a patient's logical reason and emotional data as the input data (e.g., symptoms of a disease) along with the patient's previous medical history and genetical data. Then this system can analyze those data and can predict the disease accurately by which the patient is suffering, and it will give the visualization of those data to the users. Based on the prediction disease, the system can provide the proper treatment faster than before. Accomplishing this all the medical information about the patient, can reduce redundant and expensive testing, minimize oversight of the prescribing drugs and administering medication, and able to avoid preventable deaths.

Our contribution of this paper can be summarized as follow:

- Interactive data collection scheme to get data with less effort using some of the IoT devices and sensors.

- Significant data analysis of unstructured and semi structured data to make a structured database and store those data efficiently.

- A two-way connection to using compatible data warehouse with HPC and cloud synchronization which can store the medical data for the future.
- This system can deal with symptoms of the disease, emotional data of the patient as well as the historical medical data and genetical data about the patients.
- A prediction model which can predict both cost calculation and disease prediction in a faster way.

1.7 Organization of this book

We organized this book as In Chapter II presents some related works, Chapter III describes the healthcare system architecture, Chapter IV review about brief system architecture of the system. The experimental results and discussions are presented in Chapter V, and the last Chapter VI is the conclusion.

1.8 Conclusion

In future, healthcare research will be broader, no doubt. The National Health Reform describes a range of projects and developmental objectives that are challenges for the future. There are many continuing works which are related to recognizing and to measure the output of health care services. And in future, researchers will identify those problems and solve those problems by research about healthcare. It will study the effectiveness of a home-based intervention for people caring for a family member, and review about an occupational therapy home intervention for standardizing the lifestyle and improve the life quality which the user lead. In future, for the patient hospitals will be no longer far, when our goal is to include a concept like functioning, satisfaction, painless and distress in assessment. Healthcare system research can improve current methods which will find a new technology and which will be more dependable. The healthcare services research can provide a better design by analyzing the physician and patient nature and on patient outcome. Also improve the utilization, and quality of the healthcare.

Chapter 2

Related Works

2.1 Introduction

Nowadays, medical science with smart technologies is the most emerging area that attracts many researchers to develop smart healthcare system. In this 21^{st} century, lots of advanced technologies are introduced in medical healthcare. At present, medical system must be more interactive and intelligent then traditional healthcare to deal with large biological data sets. Advance IoT based technology with the interaction of Human Computer system makes the promise to serve for a better healthcare facility which can detect disease earlier and provide smart medical system.

2.2 State of the Art

In the recent past, some method and system are proposed to make healthcare service more faster and better. In this section, we have talked about those existing technologies and healthcare system which are very closely linked to ours. We have discussed about the related works to find the limitation of those technologies and introduced some advanced features which will help to make interactive healtcare system.

2.2.1 Existing Healthcare systems and their Technologies

Healthcare system is improving continuously for the last half-decade for the tenacious research of the researchers. Nowadays, Healthcare system is becoming more advanced than previous. It has lots of Internet of Things (IoT) devices and Electronic Health Records (EHR) technology to records all of the patient's medical data efficiently. But researchers still researching about the medical system with the technologies to develop some Interactive Healthcare system. BDAEH is one of the recently proposed healthcare system, which takes the patient's logical and emotional data as the input data and then can analyze those data to predict the patient's disease [14]. But, BDAEH cannot deal with the previous medical data of the patient and also cannot analyze the patient's genetical data. AIWAC is a proposed health care system that can collect data using input data devices, analyze the received data, predict the disease and can control the interaction [15]. ReTiHA is another system which can continuously monitor the patient and can take those data as the input data and then that system can tip medical advised to the patients [16]. Some researcher's research about the healthcare with the significant data analysis and provide an algorithm for prediction model with 98% accuracy and design a probabilistic data collection with the help of different IoT devices and sensors [17]. There are also have some technique to collect medical data (for Cerebral Palsy) interactively from the patient using Electrical Medical Records (EMR) [18]. Some researchers describe the big data which is the existing to the healthcare and the future opportunities about the healthcare [19], [20]. They also showed that the research about big data analyzing in healthcare is increasing faster. There is also some research about the evaluation of big data in healthcare data using data mining and machine learning approach [21]. Some of the systems were proposed which can collect data from EHR and then use machine learning algorithm that can able to analyze those data [22]. Another Healthcare decision support system is proposed that can analyze the medical data using the help of neural network and other data mining algorithm such as Naive Bayes and C4.5 [23]. Some of the researchers use the different machine learning algorithm such as boosted trees, random forest, and neural nets. They also implemented more intelligible models such as logistic regression, naive-Bayes, and single decision trees for a case study of the hospitals [24].

Based on those case study they showed which algorithm is very efficient to predict the disease of the patient. There are also have some research about big data in medical research [25], [26].

2.2.2 Big Data in Healthcare

About the big data there are lots of research already exists which were done by the medical researchers. Some researchers were an exploration of the big data solution with the cloud computing [27], [28] and also the infrastructure of the cloud computing with big data [29], [30], [31]. There are some proposed prediction models with significant data analysis which can use any clinical or healthcare system for any types of clinical situations [32]. Some of the researchers narrate the challenges of big data and describe the systematic framework to analyze the big data analytics [33], [8], [34]. There are also some researchers about the missing data from a database, and then the researchers developed a system of Markov model which can solve the missing data [35]. Now about the medical image, to detect the information from the medical images (such as X-rays, Ultrasonography, CT scan and some MRI), there are some technologies available [36], [37]. In the medical research, some of the researchers explored the challenges of the Healthcare system like as medical image analysis, physiological signal processing and genomic data processing [38].

2.2.3 Structured Data for Healthcare

Medical data are collected from the patient continuously which will lead to store a significant amount of data and could be difficult to handle using the existing technology. The reason is those collected data are unstructured, semi structured or sometimes may be structured. Before we analyze those medical data, we need to use the structured database to avoid some unnecessary errors which can be leading to using unstructured, semi-structured dataset. MapReduce is the perfect to convert the open, semi-structured dataset to the structured dataset. MapReduce can deal with the massive amount of unstructured, semi-structured data and can turn those data to structured productively and a faster approach [39], [40], [41], [42] also MapReduce can handle the error very carefully. At the very recent researcher's research about the extended version of MapReduce as i2MapReduce for the iterative computation [43] for a significant amount of data. Some research about the task level adaptive MapReduce framework which extends MapReduce architecture by redesigning the Map and Reduce task [44]. There are also have some research about the MapReduce, and they proposed a routing algorithm known as a joint scheduler which can improve the throughput and delay performance according to Hadoop traditional fair scheduler [45].

2.2.4 Emotion Detection

Now the emotion detection, Emotions are one of the most important factors to predict the disease of a patient. Emotion is a complicated process with the various ingredients, including intuition, facial expression, cognitional reactions, thoughts or reflection, and behaviors. For the last few years, researchers were researching about the emotion detection problem, and they were fortunate to detect the emotion of a person. Many scientists investigate a lot to identify the emotion from image, video and using some wireless sensors. The Emotion API is one of the research about the emotion detection [46]. Emotion API can detect the emotional state of a person from the images and videos, and it can track out anger, happiness, sadness, neutral, fear and surprises. There are some researchers about Emotion Detection and Recognition using HRV Features [47], [48]. The heart rate variability (HRV) features extracted from photoplethysmograph (PPG) signal, which obtained from a cost-effective PPG device for detecting and recognizing the emotions based on the physiological signals using SVM. Other researchers study about, emotion detection from speech or vocal. There are also have some research about the emotion detection [49], [50]. Recently MIT researcher Dina Katabi and her team developed a new technology which can sense a human emotion via wireless signals. It can sense the emotion of a human body by transmitting RF signals and then analyzed to detect the emotion, and it can also catch the individual heartbeats from the wireless signal [1].

2.2.4.1 Heartbeat and Breathe

There are also have some sensors which can collect the medical data of a patient in an interactive way and can monitor the patient continuously [51], [2]. Vital radio can detect the heartbeat and breathe almost 99% accurately; this device can work correctly if the device is in the different room or 8 meters away from the users [52]. Vital radio can detect multiple user's heartbeats and breathe at the same time.

2.2.5 Healthcare at Present

At present, researchers continue their research about healthcare system to make an advanced interactive healthcare system. Where the system will be user-friendly and for the patient hospitals won't be the burden anymore. In this paper, we have implemented an Interactive Healthcare system design which will be helpful to all types of users.

This paper is an extended version of our previous work IHEMHA[53], a healthcare system that can deal with patient logical reasons, emotional state, previous medical history and the patient's genetic information and that was presented in a conference named ICIEV & ISCMHT 2017 held in JAPAN.

2.3 Conclusion

In this book, we introduce a smart and interactive healthcare system which can deal with large number of medical data and predict a patient's disease earlier to provide a faster service and reduce the medical cost.

Chapter 3

Architecture of Proposed system

3.1 Introduction

Our proposed interactive healthcare system architecture has focused on three phases which were shown in Figure 3.1: the data collection with the unique interaction of smart biosensor devices, analyze that massive volume of data of individual patients as well as hospitals for future disease prediction and find the pattern of similar patients. These three phases are execute in three different layer called primary layer, operating layer and application layer. In this Chapter, we will discuss the architecture of those three layer

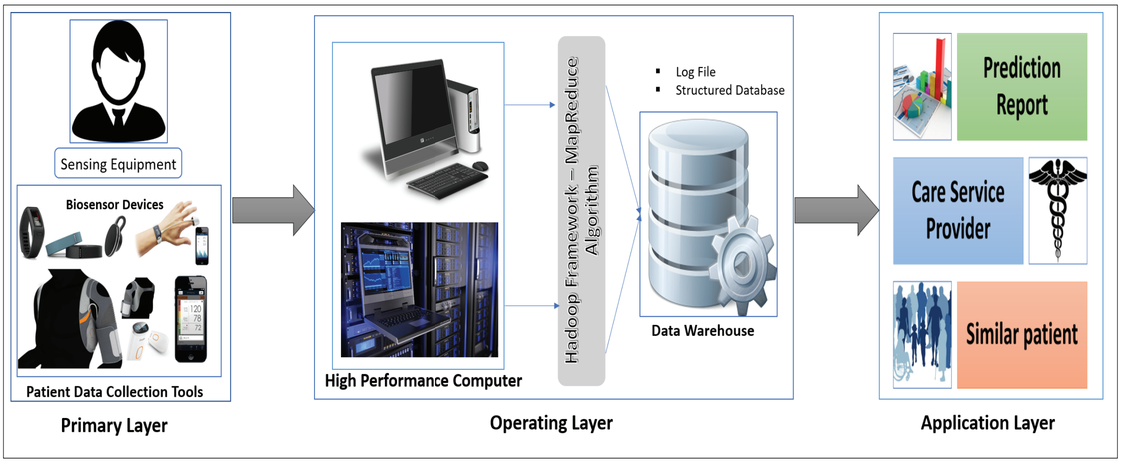

Figure 3.1: Three phases of interactive healthcare: Primary layer, Operating layer, and Application Layer

and formulate the architecture of the cloud based server for the system.

3.2 System Architecture

The system will compute the biomedical data, and emotional condition of the particular patient through smart biosensor devices and other data such as patient's medical history will be collected from hospital server. Those data will be sent to the High-Performance Computer server for primary analysis. In our system, the data collecting phase in HPC is called primary layer of data analyze scheme where all unstructured data are being gathered. The next layer is operating layer where some programming model such as Hadoop MapReduce framework will be applicable for creating a structured database with some log file and metadata. Those metadata and database will make a good user interaction in data management. After creating the log file and database, those will be integrated and sent to another high-performance server which we called a data warehouse. The log file is used as metadata for providing fast searching capability and analysis. In application layer, data warehouse sends the database to the data center for data storage which is located in the cloud. Cloud gives the facility of fast computing to diagnose a disease and predict the future disease and enables many more facilities.

3.3 Problem Formulation

A cloud based healthcare environment enables the access to store information in continuous time domain and further analysis through fast computing capabilities. High performance computer collects the healthcare raw data and manage those data for processing to explore the most valuable information. In traditional healthcare system, hospitals are the main source of generating healthcare data. We consider hospital along with individual patient as they are the major source of proposed interactive healthcare environment. Consider an interactive healthcare system with p number of patient in a set $P = \{P_1, P_2, ...P_p\}$ where $p \in P$. There are some patients who consult with doctor but doesn't need to stay overnight in the hospital. Those patients are continuously monitor through many biosensor devices. Some patient consults with doctor but need stay overnight in the hospital. We consider all patient in the set P. The major source of generating biomedical big data is the hospital. We consider h number of hospital are in a set $H = \{H_1, H_2.......H_h\}$ where $h \in H$. It is assumed that n number of department are associate with each hospital in a set $Dp = \{Dp_1, Dp_2..........Dp_n\}$ where $n \in Dp$. Let, there are d number of doctor who are working on a department in a hospital. We consider the set D^k_{ij}, where $j = \{1, 2.......d\}$ in the i^{th} department of k^{th} hospital, where $\forall i \in Dp$ and $\forall k \in H$. For example, D^5_{14} refers the 4^{th} doctor in the 1^{st} department of 5^{th} hospital. As we analyze all the big data in common HPC server thus the patient and doctors set would be $P^k_{ij} = \{P^k_{1j} \cup P^k_{2j} \cup P^k_{3j} \cup ...... \cup P^k_{nj}\}$ and $D^k_{ij} = \{D^k_{1j} \cup D^k_{2j} \cup D^k_{3j} \cup ..... \cup D^k_{nj}\}$ respectively. In the proposed model, it is assumed that data are gathered in HPC in continuous time domain thus we collected the data in a temporal time frame. Let $T = \{T_1, T_2......T_t\}$ be the set of continuous time frame where all the biomedical data are gathered. The HPC will do a linear regression over the data consider in Y axis and time in X axis for the 1^{st} prediction analysis. Then send the analyzed result to the cloud data center. HPC also sends the biomedical data to the cloud in a time interval i, where i refers the change of biological data over time. All data center $DC = \{DC_1, DC_2.......DC_\mu\}$ where $\mu \in DC$ relate to data warehouse, $WH = \{WH_1, WH_2........WH_m\}$ where $m \in WH$. The future prediction analysis will be done in the cloud server.

3.4 Conclusion

In this chapter, we describe our proposed architecture and three major portion of our proposed model. We also discuss about the problem formulation of our proposed model.

Chapter 4

Brief System Architecture

4.1 Introduction

The new healthcare plan is to provide an interactive environment to its user and some better medical facilities with important data analysis and improve the medical healthcare quality. In this regard, the exploration of most valuable information from extensive biological raw data and interactively collecting those data is the most challenging part of the new healthcare system. An interaction between human and computer is essential to get more biological data so that system can explore more valuable information. Another big challange is to store the massive volume of data for future use so that system can explore the maximum accurate result. In this chapter, we will introduce the new healthcare architecture and the functionality of each layer.

4.2 Primary Layer

The primary layer of the proposed architecture is interactive data collection scheme which helps to explore the most crucial information and discover the valuable knowledge. A human computer interaction is essential for collecting healthcare data which can play very vital role to provide sophisticated outcomes. Interactive data collection scheme will increase the efficiency of the system performance using IoT based sensor device and some other biosensor and sensor less devices.

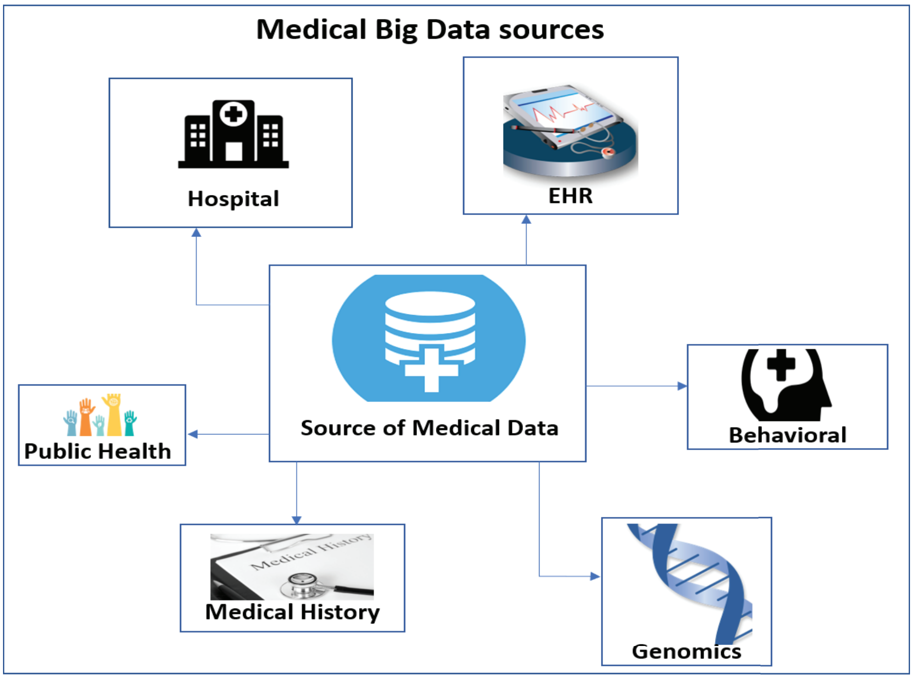

Figure 4.1: Big data source of medical healthcare.

4.2.1 Data Collection

Medical healthcare produces a massive volume of big data from different sources shown in Figure 4.1. Our proposed data collection scheme focuses on data generated by the individual patient and overall data generated by a hospital. An interaction between patient and smart device can provide a better outcome to the new healthcare system and explore more information to diagnose a disease. Nowadays, a lot of biosensors, motion sensor, wearable devices including smartphone, smartwatch are developing to measure a patient's heart rate, temperature, breath rate, ECG, back pain level and many other medical parameters. Those devices can compute a user's biological condition and transmit the result to any server. Another important fact related to health is patient's emotional state. We adopt an advance emotion-detecting technology known as EQ radio which is used to find out a user's emotional state. Patient's medical history will be collected from hospital server. Doctor's note is collected from e-prescription. Electronic Health Record (EHR) also provides many more data on an individual patient. Another significant source of data is medical clinic and hospital. Those hospital's data are being gathered from medical hospital server. A cloud-based environment can provide some social networking facilities as well as secure storage and high computing capabilities. We make interaction with all patients and doctors through cloud-based social networking to collect similar patient's public health status as well as familial genetic syndromes. Altogether all those data are sent to a high-performance computer server.

4.2.2 Data Acquisition

Our major big data source is patient and hospital. A patient generate data through the visiting frequency to a doctor in a hospital or some wearable device. Big data source of a hospital is their doctor and patient information, medical billing information and many biomedical research and medical test result. In our proposed system, we analyze the data generate by a patient to provide a better outcome in hospital and make a good interaction between doctor and patient.

Theorem 1. *An individual patient generates at least* $\sigma = \left(\frac{v}{t} * D * Dp * w\right) + g$ *amount of data in a visiting frequency to a hospital.*

Proof. Let, a patient generates w amount of data through some wearable device. As biomedical factor of a human change in time domain thus we take the latest data of wearable device. The frequency visit of a patient P_{ij}^{k} is v in time interval t to the i^{th} department of k^{th} hospital, where $\forall i \in Dp$ and $\forall k \in H$. Thus, a patient generates v/t amount of data in time interval t. There is some doctor's note which are being generated when a patient consults a doctor D and some medical test and patient information which are kept in the department's server. As we take the union set of doctors and department represent respectively as $D^k_{ij} = \{D^k_{1j} \cup D^k_{2j} \cup D^k_{3j} \cup .... \cup D^k_{nj}\}$ and $Dp_i = \{Dp_1 \cup Dp_2 \cup .... \cup Dp_n\}$. Thus, a patient generates $\big(\frac{v}{t} * D * Dp * w\big)$ amount of data in a hospital. The genomic sequence g of a patient changes very slowly thus we take that genomic sequence is constant. So, an individual patient generated σ amount of data which is shown in Equation-4.1.

$$\sigma = \left(\frac{v}{t} * D * Dp * w\right) + g \tag{4.1}$$

It is noted that the data generation of a patient increases very sharply when the visiting frequency is increased to a hospital.

Theorem 2. *Data generate by a hospital for all patient is at least* $\beta = \sum_{i=1}^{p} \alpha_i$.

Proof. A hospital consist of various departments. Those departments generate a lot of medical data. Two types of patients are generating data for a hospital. 1'st type is come to consult with doctor, but they do not stay overnight and their generation of data is at least $\big(\frac{v}{t} * D * Dp * w\big) + g$ prove in Theorem 1. The other type of patient stays overnight in hospital and generate more data. In a hospital data consist of both text data such as doctor's note and some medical test report and image data such as X-ray and signal data for example ECG or heartrate. For the patient who stay overnight in hospital, consider s, t and u megabytes be the size of each text, image and signal data respectively. The data generate by a patient in a hospital is f(TD), f(ID), and f(SD) where TD refers for text data, ID means image data and SD is signal data. A hospital generate $\Phi(P)$ amount of data for a patient, which can be expressed by Equation-4.2.

$$\Phi(P) = \sum f(TD) * s + \sum f(ID) * t + \sum f(SD) * u \tag{4.2}$$

Total medical data generation for a patient is α shown in Equation-4.3.

$$\alpha = \sigma * \Phi(P) \tag{4.3}$$

Thus, the data generate by a hospital for all patient is at least β in Equation-4.4.

$$\beta = \sum_{i=1}^{p} \alpha_i \tag{4.4}$$

4.3 Operating Layer

In our proposed interactive health care system, the massive amount of healthcare data needs to be processed and analyzed, and it requires to provide an interactive storage system which will be done in operating layer. HPC collects all unstructured and semi-structured healthcare data from different sources. To get a structured distributed file management system (SDFM), HPC applies Hadoop MapReduce framework over the unstructured or semi-structured data. Then send the SDFM to a data warehouse which makes and interaction between operating layer and application layer.

4.3.1 Hadoop MapReduce Framework

In data collection phase, HPC collects healthcare data in a continuous time domain. The processing of those extensive data is a most challenging task for getting a better output. Hadoop provides HDFS to store those data in a structured database. MapReduce algorithm enables to make the structured distributed database from raw unstructured and semi-structured data. MapReduce has two primary class: the mapper class and the reduce class. Mapper class take the data record as input and put a key with a corresponding value to the data record. Then Map function shuffle those record, sort them and send to reducer class. Reducer class marge the same keyed data as per the algorithm instruction and provide the structured database which store in HDFS. This MapReduce algorithm architecture is shown in Figure 4.2.

For applying the MapReduce framework, let us consider HPC provides θ number of active server show in a set $S = \{S_1, S_2......S_\theta\}$. These server set is used for parallel

Algorithm 1 *MapReduce Framework*

INPUT: *Unstructured or semi-structured data set.*

OUTPUT: *Create a structure data set.*

Mapper function to generate the pair of key and values

```
Input the data set and take it in InData.
for each line column and each row: do
  Data = Strip and split the line.
  if number of column in Data! = length Data.head: then
    Skip that row from Data.
  else
    Store those data generating a pair of key and values.
    Go to the next row in Data.
  end if
end for
```

Reducer function to generate the structure data

```
Set oldKey = Null.
for each line column and each row: do
  Data = Strip and split the line.
  if number of column in Data! = length Data.head: then
    Skip that row from Data.
  else if any value from Data == 0 then
    Skip that row from Data.
  else
    Store all the data in database as structure data.
  end if
end for
```

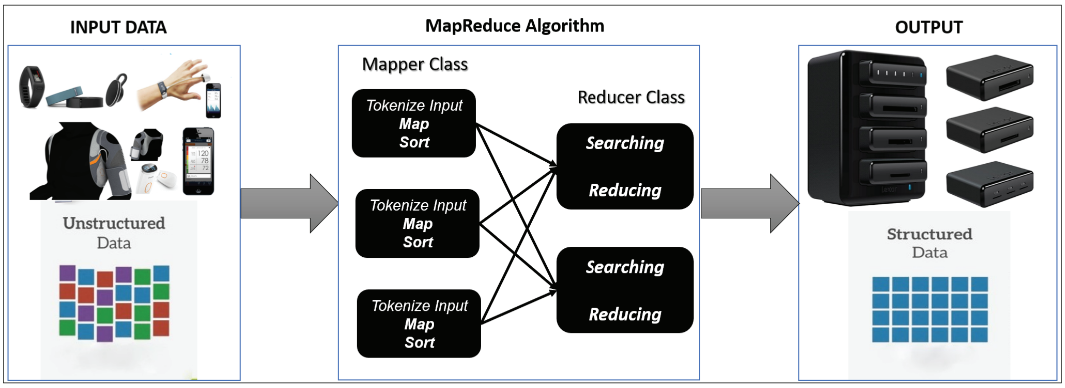

Figure 4.2: MapReduce Framework.

computation in operating layer for MapReduce framework. From the Equation-4.4, we need to analyze β amount of data. Each server receives the input dataset and make multiple partition for parallel distribution and computation. Let, the input dataset is divided into k number of partition or chunks. Thus, each partition size will be C, which is shown in Equation-4.5.

$$C = \left\lfloor \frac{\beta}{k} \right\rfloor \tag{4.5}$$

So, C amount of data quantity is distributed in each partition blocks of each server for analyzing and send to the set of Mapper class M where $M = \{M_1, M_2.......M_m\}$and m be the number of map function. The map() then tokenized all the input record with key and value and send to reducer class. Let r be the number of reducer function in a set of reducer class R where $R = \{R_1, R_2.......R_r\}$and $R \leq M$. reducer() receiver the mapped data and shuffle and sort them and provide a structured database which send to the data warehouse with all metadata and log files. In algorithm 1, there we show the MapReduce function for our structured system.

4.3.2 Architecture of Data Warehouse

Our proposed Interactive Healthcare system holds the structured clinical database in the data warehouse and enables more functionality. After analysis of data in HPC and stored them in HDFS, the database with metadata are sent to the data warehouse. Data warehouse keeps the clinical database with the modification of metadata and sends the database to the cloud server for fast computation facility with higher accuracy with privacy. The data warehouse provides a two-way interaction between HPC and Cloud server. The primary purpose of a data warehouse is to keep track of all the data in the database with metadata along with log files. It will make an interactive environment for the doctor, patient and system developer. The user can search and get the information quickly from the data warehouse, even if they are not familiar with any types programming languages. This warehouse can able to take the user question and able to convert the question as the query language. Thus, it provides the fast searching capabilities. We will apply natural language processing (NLP) to turn the question to query. The architecture of data warehouse holds two unit: metadata with log files and structured clinical database. Nodes are work under metadata to do faster search and access the clinical database. For the system developer, it is essential to provide a developing interactive environment and provide a facility to omit the complex query. Thus, metadata offers the exploratory information extraction through its log file system and distributed database. The top unit of metadata holds the knowledge graph which maps with information base and log file. In this unit, we apply NLP to extract the information of a given query very interactively. The log file contains some intent keyword such as blood pressure, respiratory rate, etc. and the knowledge of database table attribute and the relationship of every table to one another. The other part of a data warehouse is a clinical database which makes the communication between the cloud server and HPC. When a user wants to search something, data warehouse takes the question and convert into the query with the convenient word and make a map with information base and knowledge graph and send to the cluster node

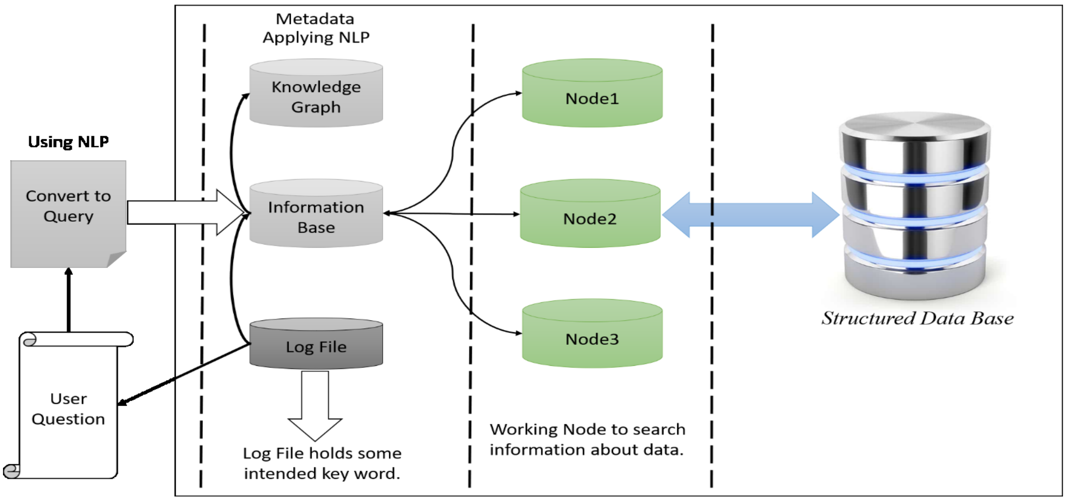

Figure 4.3: The Data Warehouse Architecture.

for parallel computing to provide the fast searching facility. In Figure 4.3, we showed the architecture of the data warehouse for an Interactive Healthcare system.

4.4 Application Layer

Application layer starts from cloud server where all users of the system need to create an account and will provide their necessary information. This is necessary to map all the users to their family member so that the system gets the genetic information of all users. Prediction analysis and patient similarity will provide a better healthcare service which will be computed in this phase.

4.4.1 Diagnosis procedure in healthcare

To understand a patient condition doctor's need to know the essential details of a patient. The primary attribute of a patient is shown in Table 4.1.

In medical diagnosis, the functional status of healthy or normal body condition is

Table 4.1: Primary Attribute of a patient

Parameters	Value
Age	Infant, Child, Adult, Old
Gender	Male, Female
Weight	Normal, Overweight, Underweight
Occupation	Mental Work, Heavy Physical Work, Light Physical Work
Physical Fitness	BMI
Smoking	Chain Smoker, Occasional Smoker, does not Smoke
Alcohol	Regular, Occasional, Never
Some Vital Sign (Temperature, Pulse Rate, BP)	Measure by observation

known as“Physiology.” The disorder of physiology is “Pathophysiology” that refers to the affected body structure or a sick patient. This happens when some other cause is applied to the physiology function. The medical term of those causes or set of causes is called “Etiology.” Several causes affect a healthy body. A disease can be caused by some factors that are genetic, environmental and epigenetic. Genetic cause means DNA mutation and factors related to DNA, environmental purpose is the severity of a disease in a geography, and epigenetic factors are the heritable changes in gene expression such as hypertension, diabetics, etc. When a patient visits a doctor, he analyzes those factors and then built some medical features. Medical features are identified by some sign, symptoms, and observation of medical test and history. The sign refers to behavioral feature for which we adopt EQ radio technology to measure user’s emotion. Symptoms

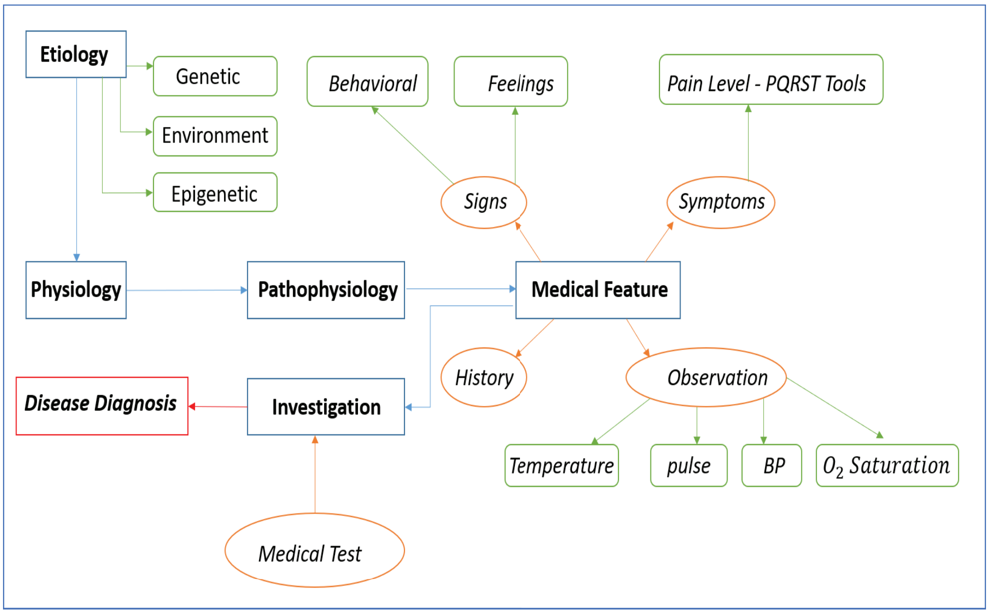

Figure 4.4: Medical Disease Diagnosis Procedure.

refers to pain level measure by PQRST pain assessment method. Medical test measures observation and another is the medical history. There are some routine tests for the consideration shown in Table 4.2. We collect that information from the smart device and medical server. After getting that information then a doctor investigates on some other features such as Ultrasound, Endoscopy for the accurate diagnosis. The diagnosis procedure is shown in Figure 4.4.

4.4.2 Prediction Model Pipeline

Prediction analysis provides the early detection of a disease which can minimize the cost of hospitalization and reduce the mortality. A proper estimation of early disease prediction can provide an accurate diagnosis and fast treatment facility to a patient. Thus, we build a prediction model pipeline to predict a disease more accurately. The

Table 4.2: Routine test attributes

Test Parameter	Normal range
Serum Glucose	4.4-7.8 mmol/l
Serum Urea	7-20 mg/dL
Serum Bilirubin	0.3-1.0 mg/dL
Urine Glucose	0.0.8 mmol/l
Urea	2.5-7.1 mmol/l
BP	80-120 (±10)
Platelet	150-400$\times 10^9$ /L
RBC	4.5-5.5$\times 10^{12}$ /L
WBC	4-11$\times 10^9$ /L

pipeline is shown in Figure 4.5.

Prediction Target:

Prediction target is the intersection of interest set of disease and the possible set of disease. Let, $I = \{I_1, I_2, I_3, ..., I_n\}$ be the interest set of disease which is calculated by etiology analysis. $\chi = \{\chi_1, \chi_2, ..., \chi_n\}$ be the set of possible disease measured by clinical vital sign and observation tool which we collect by biosensor devices. Thus, prediction target will be shown in Equation-4.6.

$$PT = I \cup \chi \tag{4.6}$$

In the top unit of prediction modeling, we use MapReduce algorithm to analyze a patient's etiology function. Patient's genetic sequence such as DNA information is constant value will be collected from patient's cloud server. The severity of diseases in a specific geography at a time interval will be collected from cloud database, and the epigenetic feature or the genetic disease which a patient expose will collect from his medical history. All those put in map class for tokenizing with key and value and then

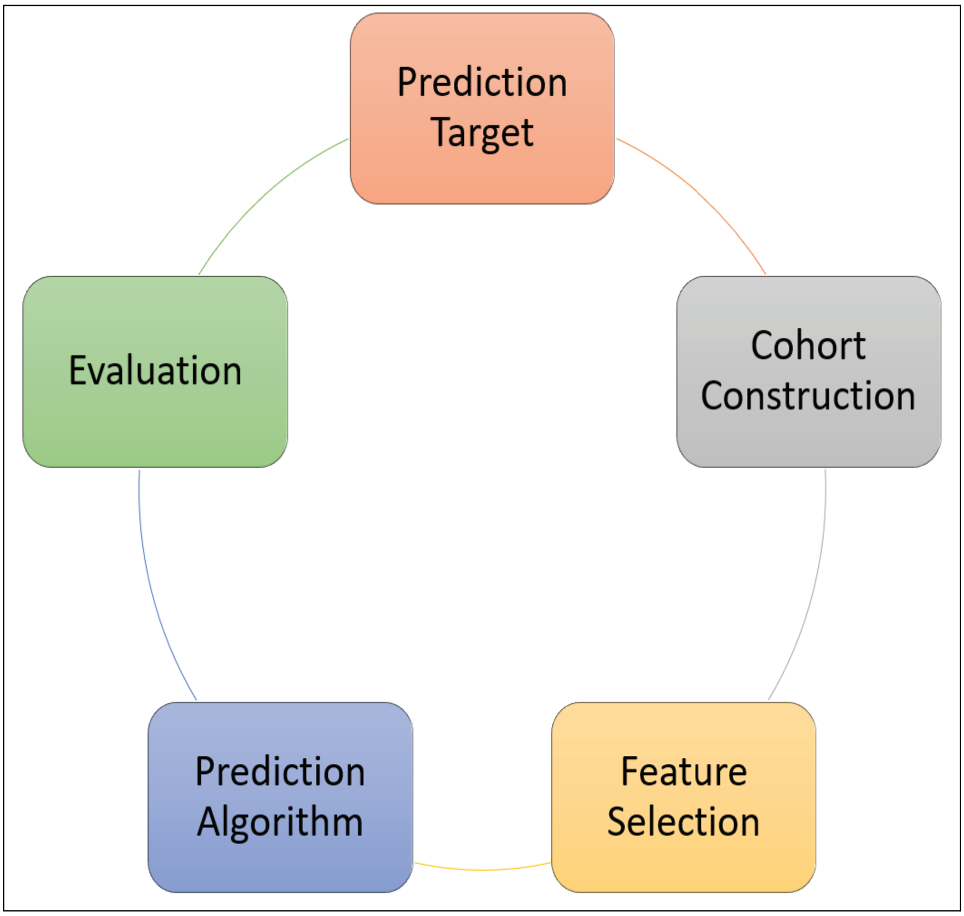

Figure 4.5: Prediction model pipeline.

send to reduce class to get the interest set of disease. We take some vital signs which we are integrated with patient's emotional state to get the possible set of disease. Vital signs are patient's heart rate, respiratory rate, some common attributes which are figure out in Table-1.

Cohort Construction:

To construct cohort, we take the study sample from the total population (all test case from the database) which we get from our database. Then identify the patient who exposes the risk factor of target prediction disease. For example, let us a target prediction of a patient is heart failure. So, in cohort construction, we take the study population of all heart failure patients. The key is to cohort study is to identify the clues of right inclusion and exclusion criteria. From cohort construction, the system will set its training dataset and test dataset and able to find the hidden evidence of disease.

Feature Selection:

The relevant clinical feature selection for the target predicted disease is another important factor in disease prediction. The complexity of prediction algorithm is depended on feature selection. A clinical feature which is related to a disease is a subset of a total number of biomedical function. Let us consider there are n number of clinical feature produced by a patient who is represented in a set $\psi = \{\psi_1, \psi_2, \psi_3..........\psi_n\}$. Among them f number of feature are responsible for a specific disease. Feature those are responsible for a disease is a subset of total feature thus we can say $f<n$.Thus, it is necessary to extract the feature among all medical attributes.

Now, if a patient generates the ∂ amount of data with n number of clinical attributes then 2^n number of subset are possible as a set of relevant features. From that, optimal subset of clinical feature must be extracted. Comparison is the most expensive operation when n is increased. Another algorithm may use for feature selection which is rank matrix where features are ranked by their information gain. But the cost of this algorithm is also higher as several operations take places for gain calculation. Thus, we apply the genetic algorithm which is shwon in algorithm 2 to calculate the fitness of the feature and eliminate those features which does not get adequate fitness.

Prediction algorithm:

Predictive model is a function of input feature of a disease and output target disease which can be mathematically represented in Equation-4.7 as

$$T = F(\psi_f) + e \tag{4.7}$$

Where T is the predicted target disease, ψ_f is the set of target feature which are extract

Algorithm 2 *Genetic algorithm*

```
INPUT: Dataset = D
Feature = f
initial-population = $\delta$
Iteration = 100

OUTPUT: Fitness value of feature.

Set crossover probability = 0.8.
Set mutation probability = 0.1.
Construct model, model = linear-regrassion(T, D).
x = matrix(model).
y = response(model).
while Iteration != 100 do
    $\delta$ = crossover($\delta$-1)
    $\delta$ = mutation($\delta$)
    f = ComputeFitness($\delta$)
end while
return f = best feature
Fitness function
ComputeFitness(input String)
Set Random Fitness = where(String = 1)
X = combine(1,x,[,Fitness])
best-fitness = linear-regrassion.fit(X,y)
```

from all clinical feature and e denotes for error. Equation-7 shows that if we know the clinical feature, past meditation with medical history, we can have identified the disease

Algorithm 3 *Prediction Model*

INPUT: *Data set.*

OUTPUT: *Create a Predicted Output.*

```
Set Factor Variables.
Take the Input data as factor (input-data).
ind = set 0.7 train data and set 0.3 test data.
if ind == 1: then
    Set the data as train-data.
else
    Set the data as test-data.
end if
Create Model.
mdl = Model(T, train-data)
pdict = predict(mdl, test-data)
```

which a patient is suffering for. This equation also shows that, the prediction model is either regression problem or classification problem. If a patient wants to predict the cost of disease which make the impact on his biological condition, then T is continuous and it's a regression problem. To predict regression problem, popular data mining technique is linear regression or generalize additive model. In classification problem, there T is categorical which denotes whether a patient suffer a disease or not. Some well-known technique for predict classification problems are Decision Tree, Naive Bayes, Artificial Neural Network, Random forest, support vector machine. To predict a disease, a simple prediction approach is shown in algorithm 3.

Evaluation Criteria:

When we construct the prediction model, we applied it to a test data set to predict the class of unseen data. So, the performance of the model should be measured so that the model gets unbiased establishment from the generation of errors. To identify a class level and the model's accuracy, the confusion matrix is a useful tool which is shown in Table 4.3. Here, we showed for two class level in confusion matrix, actual class and predicted class and the classifier are true positive, true negative, false positive, and false negative. The accuracy of the model is measured by the correctly identified of a classifier from the given test data set. Sensitivity and specificity refer true positive rate and true negative rate of the classifier which the model correctly identified. The following equation shows all the measurements.

Table 4.3: Confusion matrix and its classifier

		Predicted Class	
Actual Class		C1	C2
	C1	True Positives	False Negatives
	C2	False Positives	True Negatives

$$Classification Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$Kappa = \frac{(TotalAccuracy - RandomAccuracy)}{(1/RandomAccuracy)}$$

$$Sensitivity = TPR = \frac{TP}{TP + FN}$$

$$Specificity = TNR = \frac{TN}{TN + FP}$$

$$FPR = 1 - TNR$$

$$FNR = 1 - TPR$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Where TP = True Positive, FP = False Positive, TN = True Negative, FN = False Negative, The confusion matrix refers the risks and gains as well as cost and benefit in a classification model.

4.5 Conclusion

The new healthcare system is a bright spot in the medical field, as it makes the interaction between doctor and patient by analyzing the relationship between the biological condition of the body, emotion changing with the modification of health status to diagnosis a disease and previous health information. In this chapter, we tried to design an interactive healthcare system with big data analysis which will help the doctor for proper diagnosis a disease and provide the healthcare system a faster service.

Chapter 5

Experiment and Result

5.1 Introduction

Dealing with big data and evaluate the outcome of the system shows the efficiency of system. In this chapter, we will discuss the evaluation of the prediction algorithm and also shows the comparison between two datasets with applying different machine learning algorithm.

5.2 Environment Setup

We conducted our experiment with R programming language using RStudio (Version 1.0.153 - © 2009-2017 RStudio, Inc.). The pre-processing stage is carried out to establish a structured database in Linux Mint environment which is installed on virtual machine. MapReduce algorithm is implemented in Java Eclipse Integrated Development Environment. All the simulations are done on 64-bit Intel core i5 3.20 GHz machine.

5.3 Dataset

We collect two different sets of patient's data for our simulation. Dataset of fetal cardiotocogram (CTG) to measure fetal heart rate and uterine contraction are considered with less hidden value and have 21 features which are used to find the categorical outcome. The algorithm will predict a patient's condition in three categories: Normal, Suspect, and Pathologic. The diabetic's dataset represents the more hidden state and dissimilarity among the attributes with eight features are used to identify the binary outcome which predicts whether a patient has diabetics or not. Datasets are collected from Cleveland and Hungarian clinic [54] and available for public use.

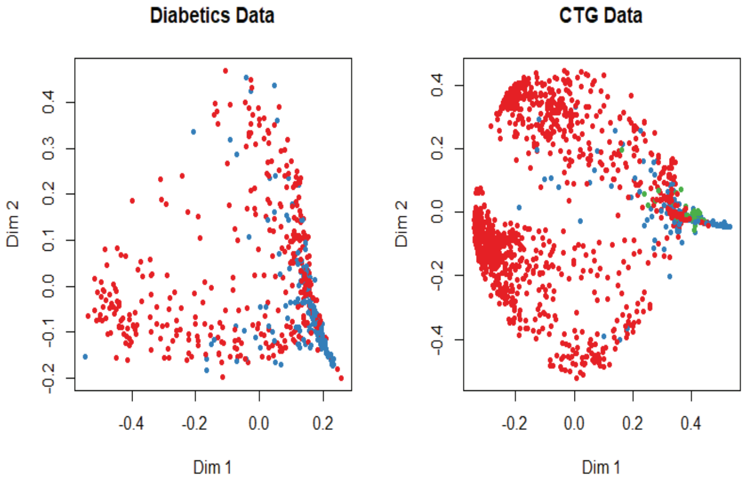

Figure 5.1: Multi-dimension scaling plot for both dataset.

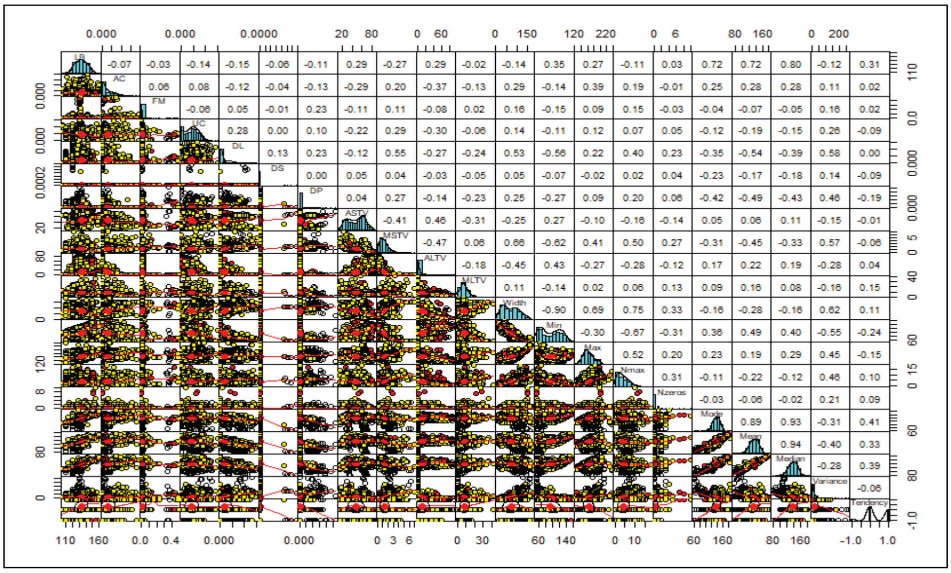

Figure 5.2: Correlation among the feature of CTG patient data.

5.4 Experiment

At the beginning state of simulation, the packages and library files are initialized for the smooth experiment. The dataset is divided into training and test data with a probability of 0.7 and 0.3 respectively. We apply several machine learning algorithms on both dataset and get different accuracy for a different algorithm. The multi-dimensional scaling plot is shown in Figure 5.1 to understand the similarity of outcome of both dataset. This plot shows that, data are more scatter for Diabetics patient than the cardiotocogram patient.

To identify the relationship between features we calculate the correlation coefficient of data. As shown in Figure 5.2 and Figure 5.3, we get the positive correlation among features of both datasets.

We incorporated genetic algorithm to eliminate the irrelevant feature. The fitness of feature over a generation of both datasets is shown in Figure 5.4. In genetic algorithm, we

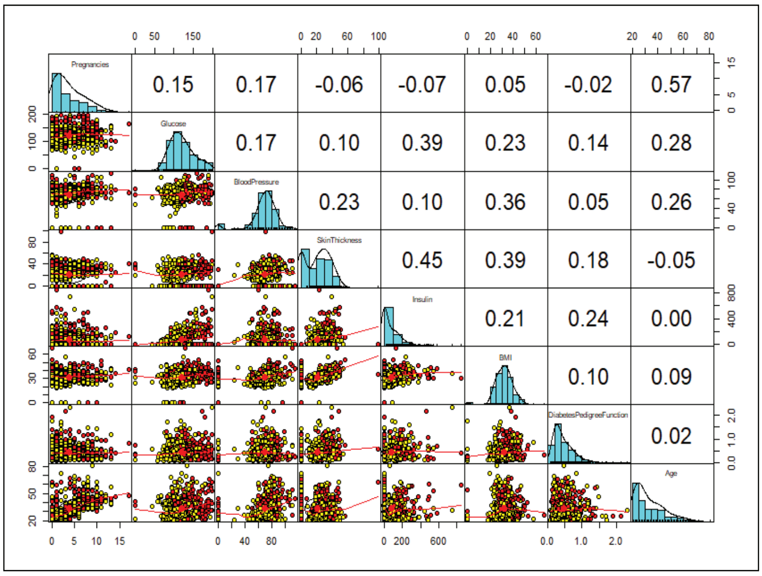

Figure 5.3: Correlation among the feature of diabetic's patient data

use the default mutation and crossover probability which is 0.1 and 0.8 respectively. After extracting features, we apply different classifier such as random forest, support vector machine, artificial neural network, C5.0 and Naive Bayes algorithm to predict a patient's disease. We also apply those classifiers before we implement the genetic algorithm to compare the accuracy and performance.

5.4.1 Performance Evaluation for Diabetics Prediction

In diabetic dataset classification, we predict the binary outcome where 0 denotes the patient who does not have diabetics, and 1 indicates the diabetes patient. After applying the prediction algorithm, we get different results for genetic and non-genetic algorithm approach for different classifiers. A comparison is shown in Table 5.1 and Table 5.2, Evaluation Criteria for Diabetics patient data which describe the prediction evaluation of all classifier for diabetic dataset before and after applying the genetic algorithm. Table

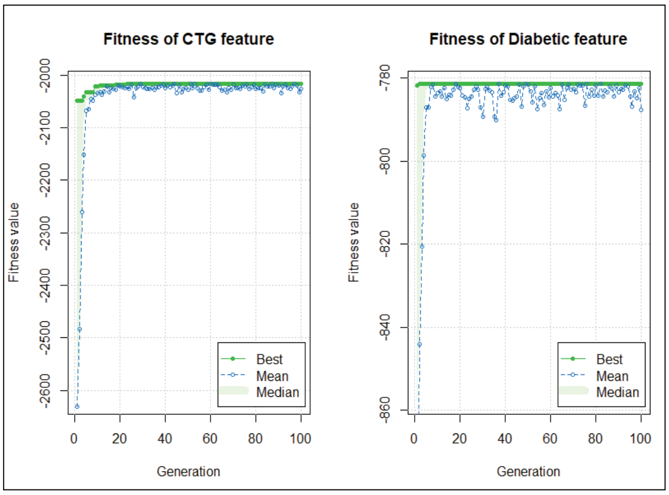

Figure 5.4: Fitness of feature for both dataset.

5.3 and Table 5.4 represents the system time which is taken to build the model.

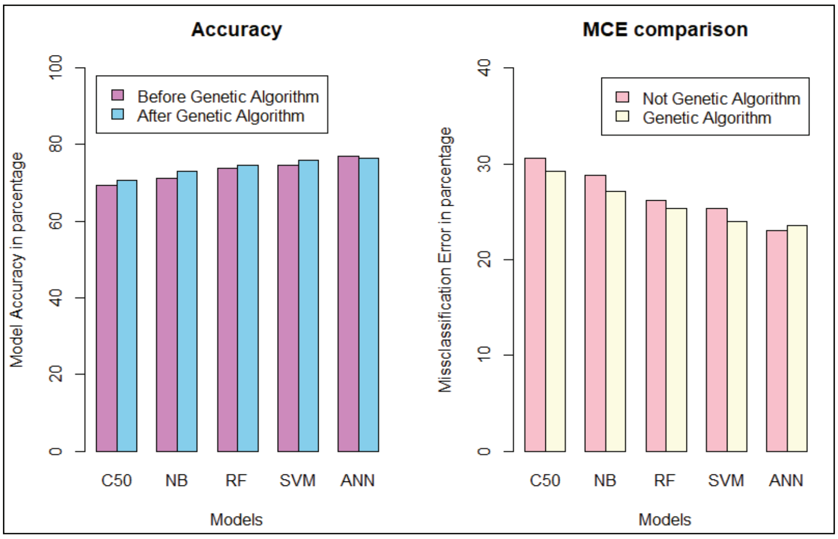

Figure 5.5: Models Accuracy and MCE for diabetic's prediction.

Table 5.1: Evaluation criteria for Diabetics prediction before implementing Genetic algorithm

	ANN	Random Forest	SVM	C5.0	Naive Bayes
Accuracy	76.92%	73.80%	74.67%	69.43%	71.17%
Kappa	50.68%	39.66%	42.60%	35.48%	37.32%
Sensitivity	81.88%	74.85%	76.36%	77.37%	76.51%
Specificity	68.67%	70.68%	70.31%	57.60%	61.25%
Precision	81.29%	88.27%	86.89%	73.10%	78.62%
Recall	81.88%	74.85%	76.36%	77.37%	76.51%
F1	81.58%	81.01%	81.29%	75.17%	77.55%
Misclassification Error	23.07%	26.20%	25.32%	30.57%	28.82%

Table 5.2: Evaluation criteria for Diabetics prediction after implementing Genetic algorithm

	ANN	Random Forest	SVM	C5.0	Naive Bayes
Accuracy	76.47%	74.67%	75.98%	70.74%	72.92%
Kappa	49.33%	43.20%	45.71%	37.48%	41.42%
Sensitivity	80.85%	77.01%	77.43%	77.46%	78.23%
Specificity	68.75%	69.12%	72.30%	59.77%	63.41%
Precision	80.01%	85.51%	87.58%	75.86%	79.31%
Recall	80.84%	77.01%	77.43%	77.46%	78.23%
F1	81.41%	81.04%	82.20%	76.65%	78.86%
Misclassification Error	23.52%	25.32%	24.01%	29.25%	27.07%

We get the highest accuracy in the artificial neural network at the diabetic dataset, but this model takes higher time to develop the model. After implementing the genetic algorithm, we can minimize the model construction time, but it slightly decreases the

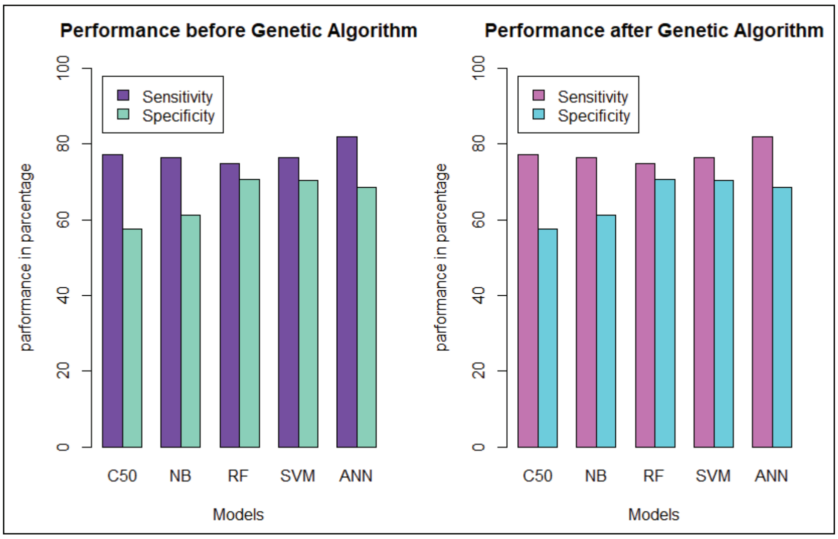

Figure 5.6: Performance evaluation for diabetics prediction.

Table 5.3: Model Construction time for Diabetics Prediction table before genetic algorithm incorporated

	ANN	Random Forest	SVM	C5.0	Naive Bayes
CPU Time	8.35	0.50	0.04	0.25	0.03
System Time	0.00	0.02	0.00	0.02	0.00
Elapsed Time	8.49	0.53	0.05	0.39	0.03

accuracy of ANN where other model's accuracy is increased. For visualizing the performance of the classifier, a graph of Sensitivity (TPR) over Specificity (FPR) is shown in Figure 5.7. The accuracy and misclassification error diagram is shown in Figure 5.5 and

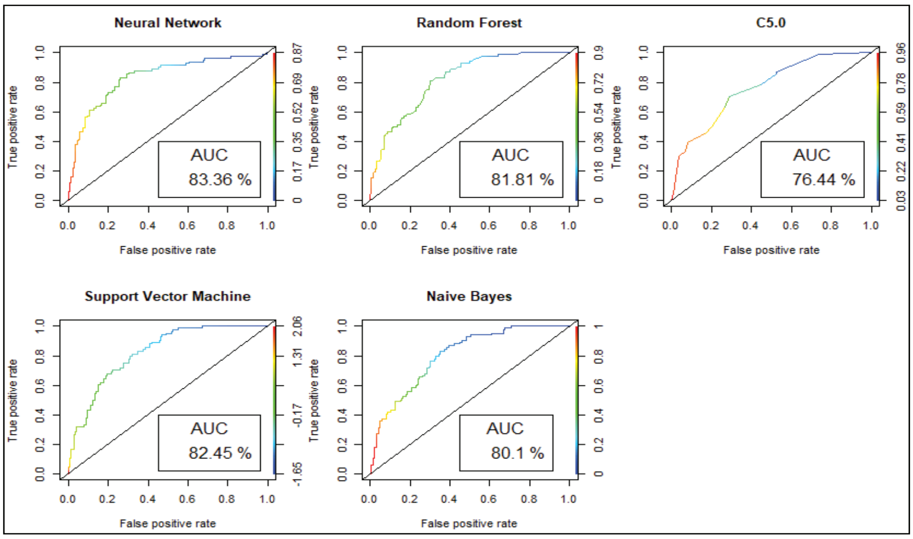

Figure 5.7: Model performance with ROC curve for Diabetics Prediction before Genetic Algorithm

Table 5.4: Model Construction time for Diabetics prediction after genetic algorithm incorporated

	ANN	Random Forest	SVM	C5.0	Naive Bayes
CPU Time	4.81	0.23	0.03	0.37	0.01
System Time	0.00	0.00	0.00	0.05	0.00
Elapsed Time	4.95	0.24	0.03	0.55	0.01

the Performance evaluation of diabetics patient shown in Figure 5.6. The ROC curve for diabetic's prediction model shows that, artificial Neural Network can predict well in both

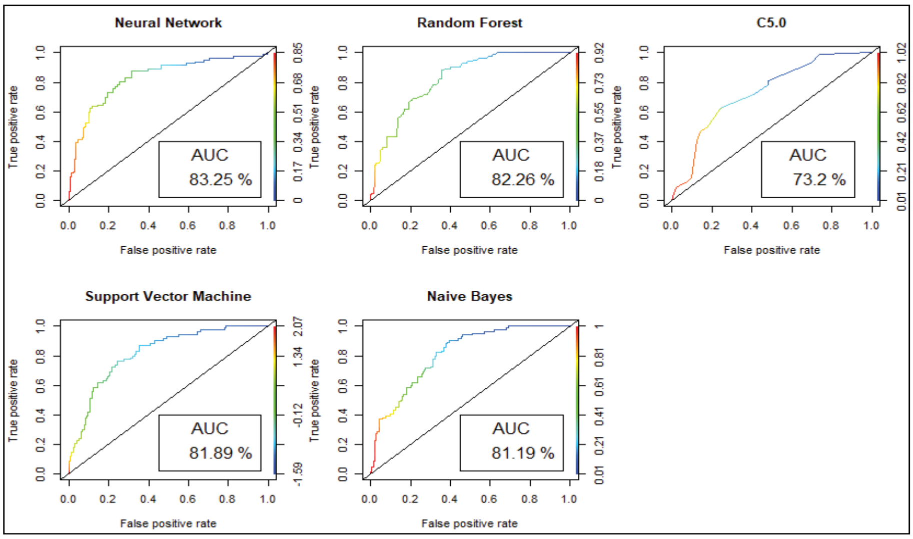

Figure 5.8: Model performance with ROC curve for Diabetics Prediction after Genetic Algorithm

scenario comparing to other model, and we get 83.36% AUC for ANN before implementing genetic algorithm. The other model's prediction is also well but the performance of C5.0 classifier is not so good for diabetic disease prediction where the dataset holds less similarity among the features.

5.4.2 Performance Evaluation for CTG Prediction

CTG dataset is used to measure fetal heart rate and uterine contraction and the algorithm predicts the outcome in three classes. The categorical outcome for CTG indicates regular patient through 1, 2 for the suspect and 3 represent the pathologic patient. In this dataset, there are enormous similarity among the cases of the feature. We get better correlation among the feature of CTG data attribute than diabetics patient which is shown in 5.2 and 5.3.

Model performance is changed for CTG dataset which has less hidden value but the more significant number of features and observation. For predicting the fetal heart rate from CTG data, we get maximum accuracy in Random Forest shown in Table 5.5 and Table 5.6. The other attribute to evaluate the model is shown in Table 5.7, Table 5.8, Table 5.9, Table-5.10. ROC curve is incorporated with the calculation of AUC in Figure 5.8 and Figure 5.9 to visualize the performance of the model with each class. Figure 5.10, Figure 5.11 and Figure 5.12 shows the statistical overview of all model with the calculation of accuracy, misclassification error and sensitivity with specificity. In this classification model, the system time is nearly similar for ANN and random forest, but random forest took maximum time comparing to other models. The model construction time is shown in Table 5.11 and Table 5.12.

Table 5.5: Evaluation Criteria for CTG patient for prediction before Genetic Algorithm

	ANN	Random Forest	SVM	C5.0	Naive Bayes
Accuracy	91.19%	93.18%	91.92%	92.55%	84.62%
Kappa	75.03%	80.98%	76.93%	79.60%	62.24%
Misclassification Error	8.80%	6.81%	8.08%	7.44%	20.24%

Table 5.6: Evaluation Criteria for CTG patient for prediction after implementing Genetic algorithm

	ANN	Random Forest	SVM	C5.0	Naive Bayes
Accuracy	89.18%	93.34%	91.13%	92.23%	84.46%
Kappa	68.48%	80.02%	73.66%	77.99%	57.69%
Misclassification Error	10.81%	6.65%	8.87%	7.76%	15.53%

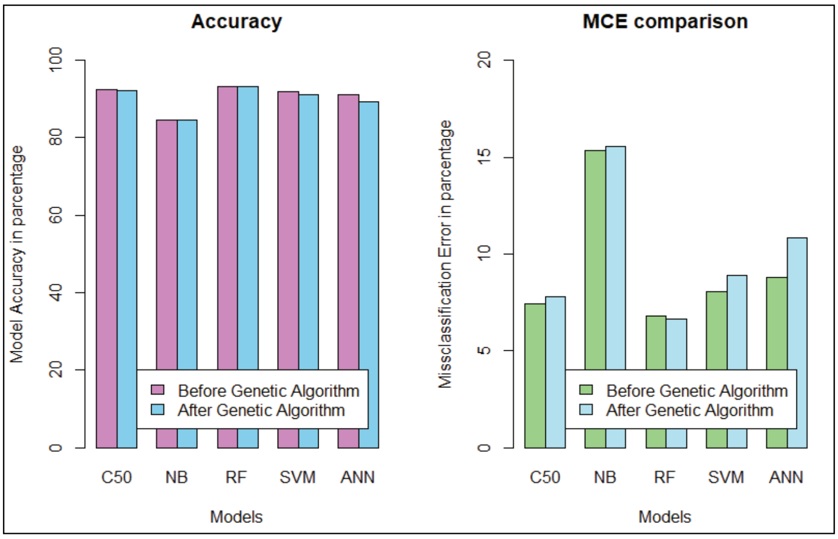

Figure 5.9: Model's Accuracy and MCE for fatal heart rate patient.

Table 5.7: Evaluation Criteria for CTG data for prediction before genetic algorithm

	Sensitivity			Specificity			Precision		
	Normal	Sus-pect	Patho-logic	Normal	Sus-pect	Patho-logic	Normal	Sus-pect	Patho-logic
ANN	94%	71%	85%	86%	95%	98%	96%	67%	79%
RF	96%	75%	88%	86%	96%	98%	95%	79%	89%
SVM	96%	73%	75%	81%	95%	99%	95%	72%	93%
C5.0	94%	85%	81%	85%	95%	99%	95%	75%	91%
NB	87%	79%	64%	88%	88%	96%	96%	51%	63%

Table 5.8: Evaluation Criteria for CTG data for prediction before genetic algorithm

	Recall			F1		
	Normal	Suspect	Pathologic	Normal	Suspect	Pathologic
ANN	93.52%	65.67%	76.36%	95.15%	57.14%	77.77%
Random Forest	97.37%	69.51%	88.89%	96.30%	74.51%	88.89%
SVM	97.57%	62.20%	75.92%	95.45%	67.11%	83.67%
C5.0	95.96%	75.61%	83.33%	95.28%	76.54%	87.38%
Naive Bayes	91.52%	53.66%	66.67%	92.07%	53.65%	63.15%

Table 5.9: Evaluation Criteria for Diabetics patient data after Genetic algorithm

	Sensitivity			Specificity			Precision		
	Normal	Suspect	Pathologic	Normal	Suspect	Pathologic	Normal	Suspect	Pathologic
ANN	93%	65%	76%	86%	92%	98%	96%	50%	79%
RF	97%	69%	90%	83%	97%	98%	95%	80%	88%
SVM	97%	62%	75%	75%	96%	99%	93%	72%	93%
C5.0	95%	75%	83%	80%	96%	99%	94%	77%	91%
NB	91%	53%	66%	73%	93%	95%	92%	53%	60%

Table 5.10: Evaluation Criteria for Diabetics patient data after Genetic algorithm

	Recall			F1		
	Normal	Suspect	Pathologic	Normal	Suspect	Pathologic
ANN	93.52%	65.67%	76.36%	95.15%	57.14%	77.77%
Random Forest	97.37%	69.51%	88.89%	96.30%	74.51%	88.89%
SVM	97.57%	62.20%	75.92%	95.45%	67.11%	83.67%
C5.0	95.96%	75.61%	83.33%	95.28%	76.54%	87.38%
Naive Bayes	91.52%	53.66%	66.67%	92.07%	53.65%	63.15%

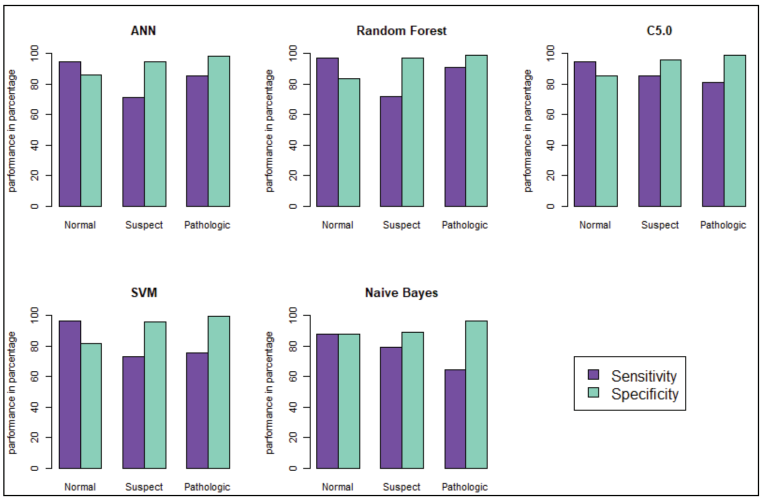

Figure 5.10: Model Performance for predicting class without implementing Genetic Algorithm.

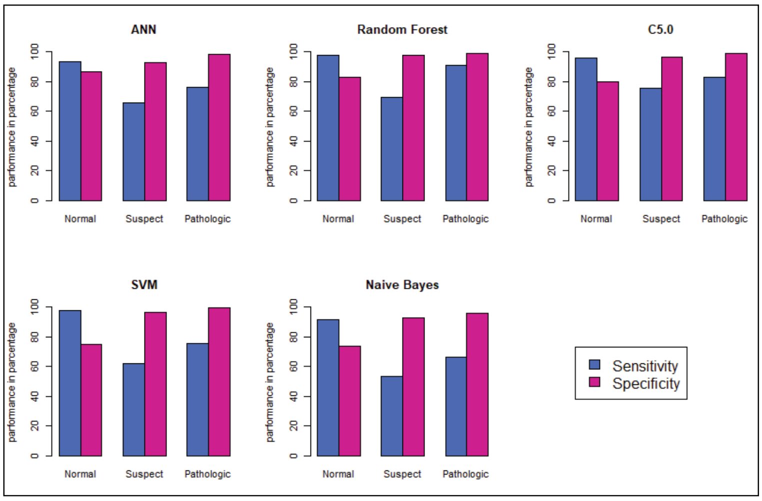

Figure 5.11: Model Performance for predicting class with implementing Genetic Algorithm.

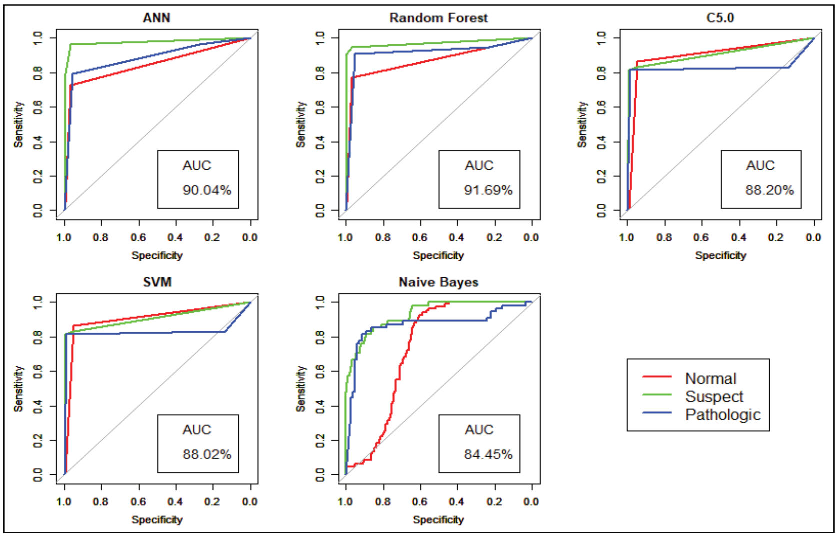

Figure 5.12: Model Performance with ROC curve before Genetic Algorithm.

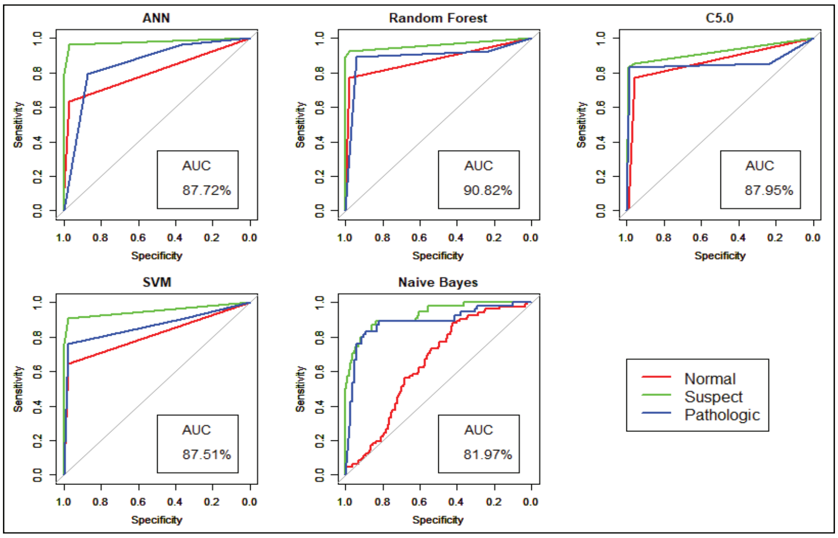

Figure 5.13: Model Performance with ROC curve after Genetic Algorithm.

Table 5.11: Model Construction time for CTG patient for prediction before Genetic Algorithm

	ANN	Random Forest	SVM	C5.0	Naive Bayes
CPU Time	1.71	1.78	0.41	0.83	0.03
System Time	0.00	0.05	0.00	0.14	0.00
Elapsed Time	1.73	1.94	0.41	1.17	0.03

Table 5.12: Model Construction time for CTG patient for prediction after genetic algorithm incorporated

	ANN	Random Forest	SVM	C5.0	Naive Bayes
CPU Time	1.09	1.60	0.17	0.07	0.00
System Time	0.00	0.00	0.00	0.00	0.00
Elapsed Time	1.16	1.64	0.20	0.06	0.00

5.5 Conclusion

From the result which we analyze for both diabetics and CTG patient, we can say that model accuracy and performance will be better if the similarity gets higher among the features of the dataset. Positive correlation among the features also make great impact in prediction of a disease.

Chapter 6

Conclusion and Future Work

6.1 Introduction

This paper describes a new healthcare system, which can deal with the large volume of patient's data along with the full overview of patient's medical conditions. Besides data collection, we have designed a probabilistic data acquisition scheme which will help to analyze the massive amount of unstructured data and those schemes are efficient for a loud environment. A data warehouse is introduced to store data and enables other functions which can make two-way interaction with HPC and cloud server. In this paper, we have implemented some prediction model algorithm on existing dataset and showed the performance of those models. We have showed a side by side comparison of some data mining technique; which are Naive Bayes, SVM, Artificial Neural Network, Random Forest, and C5.0 classifier on healthcare data.

6.2 Future Work

Our future work is to develop an analytic model for health data visualization using Augmented Reality and Virtual Reality. Signal Procession and image analysis can expose a new dimension in the healthcare industry which will allow many more features in an interactive healthcare system. Thus, dealing with signal and image data is another part of future ventures.

6.3 Conclusion

Our proposed healthcare system will improve the development of the medical healthcare which will predict a patient's disease earlier. This disease prediction system will help doctors to diagnosis a disease in a faster way and save time which may need in disease diagnosis. It will minimize the medical cost (e.g. Lab test, X-rays and some extra unnecessary medical test), maximize the medical service and provide a better outcome in medical healthcare industry.

Bibliography

[1] M. Zhao, F. Adib, and D. Katabi, “Emotion recognition using wireless signals,” in *Mobicom’16*, October 2016.

[2] M. Chen, Y. Ma, J. Song, C. F. Lai, and B. Hu, “Smart clothing: Connecting human with clouds and big data for sustainable health monitoring,” in *Mobile Netw. Appl., vol. 21, no. 5*, 2016, pp. 825–845.

[3] T. B. Murdoch and A. S. Detsky, “The inevitable application of big data to health care,” in *JAMA. 309 (13):*, April 2013, p. 1352.

[4] W. Raghupathi and V. Raghupathi, “Big data analytics in healthcare: promise and potential,” in *Health Inf. Sci. Syst., vol. 2, no. 1*, 2014, pp. 1–10.

[5] “(november. 2017), the digital universe of opportunities: Rich data and the increasing value of the internet of things.” [Online]. Available: https://www.emc.com/collateral/analyst-reports/idc-digital-universe-2014.pdf

[6] E. M. Mirkes, T. J. Coats, J. Levesley, and A. N. Gorban, “Handling missing data in large healthcare dataset: A case study of unknown trauma outcomes,” in *Computers in Biology and Medicine. 75:*, June 2016, pp. 203–216.

[7] “Ibm big data in a minute: Transforming unstructured data into better healthcare outcomes, august 26, 2014,” 2014. [Online]. Avail-

able: http://www.ibmbigdatahub.com/video/ibm-big-data-minute-transforming-unstructured-data-better-healthcare-outcomes accessed: 18. September. 2017

[8] C. McDonald, "5 big data trends in healthcare for 2017," in *https://mapr.com/blog/5-big-data-trends-healthcare-2017/ accessed: 08. October. 2017*, Feb. 13 2017, pp. 514–522.

[9] D. L. Rizzatti, "Digital data storage is undergoing mind-boggling growth," in *http://www.eetimes.com/author.asp?section_id=36&doc_id=1330462 accessed: 18. September. 2017*, Sept 9 2016.

[10] J. O'Donoghue and J. Herbert, "Data management within mhealth environments: Patient sensors, mobile devices, and databases," in *Journal of Data and Information Quality. 4 (1): 5:1-5:20*, October 2012.

[11] R. A. Hernandez, J. Dil, B. Fisher, and T. M. Green, "Visual analytics and human-computer interaction," in *interaction*, January + February 2011.

[12] . O. . Linda A. Winters-Miner, "Seven ways predictive analytics can improve healthcare," 2014. [Online]. Available: https://www.elsevier.com/connect/seven-ways-predictive-analytics-can-improve accessed: 17 September 2017

[13] Y. Lee, H. Bang, and D. J. Kim, "How to establish clinical prediction models," in *Endocrinology and Metabolism 31.1*, 2016, pp. 38–44.

[14] K. Lin, F. Xia, W. Wang, D. Tian, and J. Song, "System design for big data application in emotion-aware healthcare," in *IEEE Access, vol. 4*, 2016, pp. 6901–6909.

[15] M. Chen, Y. Zhang, Y. Li, M. M. Hassan, and A. Alamri, "Aiwac: Affective interaction through wearable computing and cloud technology," in *IEEE Wireless Commun. vol. 22, no.1*, February 2015, pp. 20–27.

[16] K. Dolui, S. Mukherjee, S. K. Datta, and V. Rajamani, "Retiha: Real time health advice and action using smart devices," in *Proc. Int. Conf. Control Instrum. Commun. Comput. Technol. (ICCICCT)*, July 2014, pp. 979–984.

[17] P. K. Sahoo, S. K. Mohapatra, and S. L. Wu, "Analyzing healthcare big data with prediction for future health condition," in *IEEE Access, vol. 4*, 2016, pp. 9786–9799.

[18] I. Raharjo, T. G. Burns, J. Venugopalan, and M. D. Wang, "Development of user-friendly and interactive data collection system for cerebral palsy," in *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), Las Vegas, NV*, 2016, pp. 406–409.

[19] J. Andreu-Perez, C. C. Y. Poon, R. D. Merrifield, S. T. C. Wong, and G. Z. Yang, "Big data for health," in *IEEE Journal of Biomedical and Health Informatics, vol. 19, no. 4*, July 2015, pp. 1193–1208.

[20] A. Das, T. Adhikary, M. Razzaque, M. Alrubaian, M. M. Hassan, Z. Uddin, and S. Biao, "Big media healthcare data processing in cloud: a collaborative resource management perspective," in *Journal:Cluster Computing*, March 2013, p. 160.

[21] B. Boukenze, H. Mousannif, and A. Haqiq, "Predictive analytics in healthcare system using data mining techniques," in *CS & IT-CSCP*, 2016, pp. 1–9.

[22] K. Ng, A. G. S. R. Steinhubl, W. F. Stewart, B. Malin, and J. Sun, "Paramo: A parallel predictive modeling platform for healthcare analytic research using electronic health records," in *J Biomed Inform. 48:*, April 2014, pp. 160–170.

[23] M. Dey and S. S. Rautaray, "Study and analysis of data mining algorithms for healthcare decision support system," in *IJCSIT,Vol. 5 (1)*, 2014, pp. 470–477.

[24] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmis-

sion," in *Proceeding KDD '15 Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,Sydney, Australia*, 2015, pp. 1721–1730.

[25] S. Ram, W. Zhang, M. Williams, and Y. Pengetnze, "Predicting asthma-related emergency department visits using big data," in *IEEE Journal of Biomedical and Health Informatics, vol. 19, no. 4*, July 2015, pp. 1216–1223.

[26] H. Wu, W. Pan, X. Xiong, and S. Xu, "Human activity recognition based on the combined svm & hmm," in *2014 IEEE International Conference on Information and Automation (ICIA), Hailar*, 2014, pp. 219–224.

[27] M. Assuncao, R. Calheiros, S. Bianchi, M. Netto, and R. Buyya, "Big data computing and clouds: Trends and future directions," in *J. Parallel Distrib. Comput.*, 2014.

[28] T. Adhikary, A. Das, M. Razzaque, M. Alrubaian, M. M. Hassan, and A. Alamri, "Quality of service aware cloud resource provisioning for social multimedia services and applications," in *Multimedia Tools and Applications, vol. 76, no. 12*, June 2017, pp. 14 485–14 509.

[29] A. Das, T. Adhikary, M. Razzaque, E. Cho, and C. Hong, "A qos and profit aware cloud confederation model for iaas service providers," in *Proceedings of the 8th International Conference on Ubiquitous Information Management and Communication (ACM)*, January 2014, p. 42.

[30] T. Adhikary, A. Das, M. Razzaque, A. Almogren, M. Alrubaian, and M. M. Hassan, "Quality of service aware reliable task scheduling in vehicular cloud computing," in *Mobile Networks and Applications, vol. 21, no. 3*, June 2016, pp. 482–493.

[31] T. Adhikary, A. Das, M. Razzaque, and A. Sarkar, "Energy-efficient scheduling algorithms for data center resources in cloud computing," in *High Performance Comput-*

ing and Communications & 2013 IEEE International Conference on Embedded and Ubiquitous Computing (HPCC-EUC), 2013 IEEE 10th International Conference on IEEE, November 2013, pp. 1715–1720.

[32] Y. Lee, H. Bang, and D. J. Kim, "How to establish clinical prediction models," in *Endocrinol Metab (Seoul). 31(1):*, March 2016, pp. 38–44.

[33] H. Hu, Y. Wen, T. S. Chua, and X. Li, "Toward scalable systems for big data analytics: A technology tutorial," in *IEEE Access, vol. 2*, 2014, pp. 652–687.

[34] K. Dhamdhere, K. McCurley, M. Sundararajan, Q. Yan, and R. Nahmias, "Analyza: Exploring data with conversation," in *Intelligent User Interfaces 2017, ACM, Limassol, Cyprus*, March 2017.

[35] E. Mirkes, T.J.Coats, J.Levesley, and A.N.Gorban, "Handling missing data in large healthcare dataset: A case study of unknown trauma outcomes," in *Computers in Biology and Medicine 75*, 2016, pp. 203–216.

[36] C. M. Chen, Y. H. Chou, N. Tagawa, and Y. Do, "Computer-aided detection and diagnosis in medical imaging," in *Computational and Mathematical Methods in Medicine, Volume 2013, Article ID 790608, 2 page*, 2013.

[37] K. Bhima and A. Jagan, "Analysis of mri based brain tumor identification using segmentation technique," in *2016 International Conference on Communication and Signal Processing (ICCSP), Melmaruvathur*, 2016, pp. 2109–2113.

[38] A. Belle, R. Thiagarajan, S. M. R. Soroushmehr, F. Navidi, D. A. Beard, and K. Najarian, "Big data analytics in healthcare," in *BioMed Research International, Vol 2015, Article ID 370194, 16 page*, 2015.

[39] F. N. Afrati and J. D. Ullman, “Optimizing multiway joins in a map-reduce environment,” in *IEEE Transactions on Knowledge and Data Engineering, vol. 23, no. 9*, Sept 2011, pp. 1282–1298.

[40] J. Dean and S. Ghemawat, “Mapreduce: simplified data processing on large clusters,” in *Proceeding OSDI’04 Proceedings of the 6th conference on Symposium on Opearting Systems Design & Implementation - Volume 6, San Francisco, CA*, December 06 - 08 2004, p. 10.

[41] T. Condie, N. Conway, P. Alvaro, J. M. Hellerstein, K. Elmeleegy, and R. Sears, “Mapreduce online.”

[42] H. Lu, C. Hai-Shan, and H. Ting-Ting, “Research on hadoop cloud computing model and its applications,” in *2012 Third International Conference on Networking and Distributed Computing, Hangzhou, China*, 2012, pp. 59–63.

[43] Y. Zhang, S. Chen, Q. Wang, and G. Yu, “i2 mapreduce: Incremental mapreduce for mining evolving big data,” in *IEEE Trans. Knowl. Data Eng., vol. 27, no. 7*, Jul. 2015, pp. 1906–1919.

[44] F. Zhang, J. Cao, S. U. Khan, K. Li, and K. Hwang, “A task-level adaptive mapreduce framework for real-time streaming data in healthcare applications,” in *Future Generat. Comput. Syst., vols. 43-44*, Feb 2015, pp. 149–160.

[45] W. Wang and L. Ying, “Data locality in mapreduce: A network perspective,” in *Perform. Eval., vol. 96*, Feb. 2016, pp. 1–11.

[46] “Microsoft project oxford emotion api.” [Online]. Available: https://azure.microsoft.com/en-us/services/cognitive-services/emotion/, accessed: 27-Sep-17

[47] R. Rakshit, V. R. Reddy, and P. Deshpande, "Emotion detection and recognition using hrv features derived from photoplethysmogram signals," in *ERM4CT'16, Japan*, November 2016.

[48] D. S. Quintana, A. J. Guastella, T. Outhred, I. B. Hickie, , and A. H. Kemp, "Heart rate variability is associated with emotion recognition: direct evidence for a relationship between the autonomic nervous system and social cognition," in *International Journal of Psychophysiology, vol. 86, no. 2*, 2012, pp. 168–172.

[49] J. S. et al, "Heart rate variability: a noninvasive electrocardiographic method to measure the autonomic nervous system," in *Swiss medical weekly, vol. 134*, 2004, pp. 514–522.

[50] T. Penzel, J. W. Kantelhardt, L. Grote, J. H. Peter, and A. Bunde, "Comparison of detrended fluctuation analysis and spectral analysis for heart rate variability in sleep and sleep apnea," in *IEEE Transactions on Biomedical Engineering, vol. 50, no. 10*, Oct 2003, pp. 1143–1151.

[51] S. Hu and J. Tan, "Biologger: A wireless physiological sensing and logging system with applications in poultry science," in *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Minneapolis, MN*, 2009, pp. 4828–4831.

[52] F. Adib, H. Mao, Z. Kabelac, D. Katabi, and R. C. Miller, "Smart homes that monitor breathing and heart rate," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, ACM, 2015*, 2015, pp. 837–846.

[53] M. R. Ullah, M. A. R. Bhuiyan, and A. K. Das, "Ihemha: Interactive healthcare system design with emotion computing and medical history analysis," in *ICIEV & ISCMHT 2017,IEEE, Himeji, Japan*, September 1-3 2017.

[54] M. L. (2013), "Uci machine learning repository." [Online]. Available: http://archive.ics.uci.edu/ml, accessed: 23. September. 2017

Appendix A

List of Acronyms

ANN	Artificial Neural Network
API	Application Program Interface
AUC	Area Under The Curve
BMI	Body Mass Index
BP	Blood Pressure
CT	Computed Tomography
CTG	Cardiotocogram
DNA	Deoxyribonucleic Acid
ECG	Electrocardiogram
EHR	Electronic Health Records
EMG	Electromyography
EMR	Electrical Medical Records
FNR	False Negative Rate
FPR	False Positive Rate
HCI	Human Computer Interaction
HDFS	Hadoop Distributed File System
HPC	High Performance Computing
HRV	Heart Rate Variability
IDC	International Data Corporation
IoT	Internet of Things

MCE	Misclasification Error
MRI	Magnetic Resonance Imaging
NB	Naive Bayes
NLP	Natural Language Processing
PPG	Photoplethysmograph
PQRST	Personal Questionnaire Rapid Scaling Technique
RBC	Red Blood Cell
RF	Random Forest
ROC	Receiver Operating Characteristic
SDFM	Structured Distributed File Management System
SVM	Support Vector Machine
TNR	True Negative Rate
TPR	True Positive Rate
WBC	White Blood cell

Appendix B

List of Notations

$\in$	Epsilon
$=$	Equal
$\forall$	Forall
$\cup$	Union
P	Patient
D	Doctors
H	Hospital
WH	Warehouse
T	set of continuous time frame
D_p	Department
DC	Data Center
v	total visit of Hospital by Patient
g	Gnomic Sequence
σ	amount of data generated by each patient
β	Total Data generate by a hospital
TD	Text Data
ID	Image Data
SD	Signal Data
$\Phi(P)$	Hospital generate amount of data for a patient
α	Total medical data generation for a patient

θ	number of active server
k	number of partition or chunks
C	partition size
M	Mapper Class
R	Reducer Class
$\leq$	less then or equal
!=	cannot equal
==	Equals to
I	interest set of disease
χ	set of possible disease measured by clinical vital sign and observation tool
PT	Prediction Target
μ	means number of Data Center
ψ_f	set of target feature
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative

Appendix C

List of Publications

International Journal Papers

1. M. A. R. Bhuiyan, M. R. Ullah, A. K. Das, *"iHealthcare:* An interactive healthcare system of big data application with predictive model analysis", IEEE Transaction on Human Machine Systems. (Under Review)

International Conference Papers

2. M. R. Ullah, M. A. R. Bhuiyan, A. K. Das, *"IHEMHA:* Interactive Healthcare System Design with Emotion Computing and Medical History Analysis",ICIEV & ISCMHT 2017,IEEE, Himeji, Japan, September 1-3, 2017 (Accepted).