DESIGN MACHINE LEARNING CLASSIFIERS FOR HEALTHCARE APPLICATIONS

Thesis submitted in fulfilment of the requirements for the Degree of

DOCTOR OF PHILOSOPHY

By

Anand Kumar Srivastava

Enrolment No. 166210

Department of Computer Science Engineering and Information Technology

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY

WAKNAGHAT, SOLAN-173234, HIMACHAL PRADESH, INDIA

Dec, 2020

TABLE OF CONTENTS

SUPERVISOR'S CERTIFICATE

This is to certify that the work in the thesis entitled **"Design Machine Learning Classifiers for Healthcare Applications"** submitted by **Anand Kumar Srivastava** is a record of an original research work carried out by him under my supervision and guidance in partial fulfillment of the requirements for the award of the degree of Doctor of Philosophy in Computer Science and Engineering in the Department of Computer Science and Engineering, **Jaypee University of Information Technology, Waknaghat, INDIA.** Neither this thesis nor any part of it has been submitted for any degree or academic award elsewhere.

Date:

Dr. Yugal Kumar
Assistant Professor (Senior Grade)
Department of Computer Science & Engineering and Information Technology
Jaypee University of Information Technology
Waknaghat -173234, INDIA

Dr. Pradeep Kumar Singh
Professor
Department of Computer Science & Engineering
ABES Engineering College, Ghaziabad, UP, India

DECLARATION OF SCHOLAR

I hereby declare that the work reported in the Ph.D. thesis entitled **“Design Machine Learning Classifiers for Healthcare Applications”** submitted at **Jaypee University of Information Technology, Waknaghat, INDIA** is an authentic record of my work carried out under the supervision of **Dr. Yugal Kumar and Dr. Pradeep Kumar Singh.** I have not submitted this work elsewhere for any other degree or diploma. I am fully responsible for the contents of my Ph.D Theses.

Anand Kumar Srivastava

Department of Computer Science & Engineering and Information Technology

Jaypee University of Information Technology

Waknaghat -173234, INDIA

Date:

ACKNOWLEDGEMENTS

First of all, I would like to express my deep sense of respect and gratitude towards my guide Dr. Yugal Kumar and Dr. Pradeep Kumar Singh in the Department of Computer Science and Engineering and Information Technology, for their support throughout this research work. I want to thank them for introducing me to the field of Data Science and giving me the opportunity to work under them. Without their invaluable advice and assistance, it would not have been possible for me to complete this thesis. I am greatly indebted to them for their constant encouragement and invaluable advice in every aspect of my academic life. I consider it my good fortune to have got an opportunity to work with such wonderful persons.

I would like to express my gratitude to our Honorable Vice Chancellor Prof. (Dr.) Vinod Kumar. I would also like to thank Prof. (Dr.) Samir Dev Gupta, Dean and Head Department of CSE/IT for constant guidance, research facilities and resources to carry out my research work. I would also like to thank my doctoral committee members and other faculty members for their valuable feedback and critical reviews during presentations and time to time help.

I am also grateful for the support received from the JUIT, Waknaghat. In particular, I thank, Ravi Raina and all staff at the Department of Computer Science and Engineering and Information Technology, JUIT, Waknaghat who has been extremely helpful at numerous occasions. I thank my fellow PhD friends for their consistent help and valuable discussion. Last but not the least, I would like to thank my family my parents, for giving birth to me at the first place and supporting me spiritually throughout my life.

.

Anand Kumar Srivastava

ABSTRACT

Machine learning techniques are popular tool adopted for medical diagnosis and one of the core component of medical diagnostic system. The objective of machine learning techniques is to provide accurate and timely diagnostic results during disease diagnosis phases. Further, it also helps the physicians and medical practitioner regarding disease diagnosis. The objective of this work is to improve the diagnostic accuracy of computer aided diagnostic system. Large number of machine learning techniques are integrated in the computer aided diagnostic system for the prediction of the diseases. These machine learning techniques consider different features of disease to diagnosis the disease. It is seen that all features are not equally important in diagnostic process and irrelevant features can lead to low prediction rate. Hence in medical field, identification of irrelevant features is warm area of research. To identify the relevant features for disease prediction, attribute weighting methods are adopted. It is observed relevant features can improve the diagnostic accuracy of computer aided systems. Hence, to improve the diagnostic accuracy rate, a k harmonic mean-based attribute weighting method is developed, called KhmAW. Further, the proposed KhmAW method is integrated with SVM method, called KhmAW-SVM. In KhmAW-SVM, KhmAW method is used to identify the relevant features from dataset and SVM method is applied for diagnosis the disease. The proposed method classifies the datasets into healthy and non-healthy classes. In this work, an effort is made to design an automated diagnostic model for the diagnosis and prediction of diabetes patients. The proposed diagnostic model is designed using artificial bee colony (ABC) algorithm and deep neural network (DNN) technique, called ABC-DNN based diagnostic model. The ABC algorithm is applied to determine the relevant features for diabetes prediction and diagnosis. While, DNN technique is adopted for the prediction and diagnosis of diabetes affected patients. Many researchers address the issue of missing value in

medical data, either detect the missing value and delete the respective data instances from the dataset or adopt some default methods such as mean, median, neighbour etc., for filling the missing value. However, both methods are lacking to produce optimal results. Furthermore, outliers are also presented in data and degraded the performance of classifier. Few researchers also focus on the outlier detection in medical dataset, but it is not fully explored till date. This work also considers the two well-known problems of data i.e., i) missing value imputation, and ii) outlier. The missing value imputation issue is addressed through K-Mean^{++} based data imputation technique. This technique also validates the data through clustering and also compute the values for missing data. The outlier can be detected through an ABC based outlier detection technique. Further, the final outcome is determined using LS-SVM classifiers. Hence, this work presents a hybrid disease diagnosis framework for diabetes prediction, called hybrid diabetes prediction framework. The diabetes is also chronic disease that can affect the life of people due to high level of sugar in blood. The sugar level in blood is increased due to lack of production of insulin in human body. Large numbers of people are affected with diabetes till date and it can increase tremendously due life style behavior. Diabetes can also affect the other human organs like kidney, heart, retina and lead to failure of these organs**.** This works also presents a diabetes monitoring system to determine the risk of diabetes based on the personal health record of patients. Several rules are designed to accurate prediction of diabetes disease. The effectiveness of diabetes monitoring system is tested on a set of two hundred forty people.

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

ABC-FS	Artificial Bee Colony Optimization-Feature Selection
KHM-AW	k-harmonic mean-based attribute weighting
ABC	Artificial Bee Colony Optimization
ABC-DNN	Artificial Bee Colony Optimization-Deep Neural Network
ANFIS	Adaptive Neuro Fuzzy Inference System
ANN	Artificial Neural Network
AUC	Area Under Curve
BMI	Body Mass Index
BP	Blood Pressure
CBR	Case Based Reasoning
CDSS	Clinical Decision Support System
CL	cholesterol
CPL	cyber-physical localization
DNN	Deep Neural Network
DSS	Decision support System
DT	Decision Tree
DTLS	Datagram Transport Layer Security
FS	Feature Selection
GA	Genetic Algorithm
GA-SVM	Genetic Algorithm-Support Vector Machine
HD	Heart Disease

HDL	high-density lipoproteins
IoT	Internet of Things
KHM	k-harmonic mean
KKR	Kernel Ridge Regression
K-NN	k-nearest neighbor
LDA	Linear Discriminate Analysis
LDL	Low-Density Lipoproteins
LR	Logistic Regression
LS-SVM	least square- support vector machine
LSTM-RNN	long short-term memory with recurrent neural network
ML	Machine Learning
MLP	Multi-layer Perceptron
MLPNN	Multi-layer Perceptron Neural Network
NB	Naive Bayes
PHR	Personal Health Record
RBF-SVM	Radial Basis Function-Support Vector Machine
RF	Random Forest
RT	Random Tree
RT-PCR	Reverse transcription polymerase chain reaction
SVM	Support Vector Machine
UCI	University of California Irvine

CHAPTER 1

INTRODUCTION

1.1 Introduction

Due to advancement in health care system, life expectancy of human is increased tremendously in last two decades. But several challenges are associated with healthcare systems. These challenges are lack of medical information, diagnostic errors, inadequate data, irrelevant features, data threaten etc. In literature, several expert systems, decision diagnostic system, electronic health record system, prediction system are presented [1-3]. These systems considerably help the physicians for disease diagnosis. The term disease diagnosis refers to determine the disease through symptoms and it can be formulated as classification of medical data for decision making. Such decision-making systems requires higher computing power to process the large medical dataset. The decision capabilities of these systems depend on amount of training data. The aim of these systems is to minimize the physician errors. Such systems can enhance the decision capabilities of physician as well as improve the diagnostic accuracy. It is noticed that many computer-aided diagnostic tools are presented in literature to help the physicians. Large number of machine learning techniques is incorporated in diagnostic tools to improve the prediction rate [4-7].

On the other side, Diabetes is one of popular chronic disease that can affect the life of people throughout world. The main reason of diabetes is lack of production of insulin in human body. It can also characterize as metabolic disease having high blood sugar rate. The blood sugar level can vary due to insulin emission. Further, it can also affect the other body parts of human and it becomes worse, if it cannot treat timely. The reduced level of insulin can also communicate blood glucose in blood platelets. In turn, the chance of several other diseases like heart disease, paralysis, strokes can also increase in diabetes affected people. It is noticed that in developed countries, diabetes become well-known reason of death and it one of most common non-communicable diseases throughout world [8]. Till 2025, 300 million people will be either diabetic or pre-diabetic in world. A study reveals that fifty million people in India are either pre-diabetic or diabetic [9]. Further, this study also highlights that nearly forty-four people in India are not know, they are diabetes affected. Diabetes is a condition in which blood sugar level in human body is too high and body is not produced so much insulin to control the level of blood sugar. It is stated that diabetes can occur at any stage in human being. Clinically, diabetes can be categorized in type 1, type 2, and gestational diabetes [10]. Type 1 diabetes can be described as juvenile diabetes and it can occur in children frequently. But it can also occur in adults. In Type 1 diabetes, body is unable to produce insulin in appropriate amount and it is also destroying the cell which is responsible to produce insulin. While, Type 2 diabetes can occur any stage of human life, but it mainly occurs in fatty, middle aged and elder peoples. The gestational diabetes is occurred in women due to hormonal and other changes. Hence, diabetes can be characterized as a condition in which the body is not produced adequate insulin to convert the glucose in energy. The diabetes mellitus is referred as diabetes. It is a chronic disease associated with high levels of glucose in the blood. The main reason for the occurrence of diabetes is

i) Inadequate production of insulin through pancreas and in turn lowers blood glucose called Type 1 diabetes

ii) Inadequate sensitivity of cells to the action of insulin- Type 2 diabetes.

1.2 Motivation

To reduce the number of deaths due to diabetes, it is important that to develop new methods and techniques for the earlier detection and management of diabetes. It is noticed that late diagnosis increases the number of deaths. Hence, the integration of data mining and information technology can be utilized as suitable cutting-edge technological solution for earlier detection and diagnosis of diabetes. It is seen that all features are not equally important for decision making process. The irrelevant features can also affect the prediction accuracy of the algorithm. In literature, numerous algorithms are presented for predicting diabetes accurately. These algorithms include various conventional machine learning methods like decision tree, logistic regression, neural network etc. [11-13]. Some meta-heuristic algorithms and feature selection methods are also reported to determine the risk factors of diabetes disease [14-17]. It is also noticed that some ensemble methods are also developed for diabetes diagnosis [18-21]. Through extensive literature review, it is observed that several shortcomings are associated with diagnosis of diabetes disease.

- Prediction accuracy is one of the major factors associated with machine learning algorithm to diagnosis of diabetes disease.
- Selection of optimum diabetes features for the prediction task.
- Data imputation methods for determining the optimal value of missing data.
- Need of personal health record-based system for monitoring and diagnosis of diabetes patients.

1.3 Objectives of the Work

In past few decades, large numbers of machine learning algorithms have been developed to obtain optimum results for diabetes prediction. Hence, the objectives of this thesis work are listed as

- Deployment of Automatic Diagnostic model for the diagnostic and prediction of diabetes disease using Real data.
- Developing an efficient Machine Learning technique to improve the accuracy rate of existing technique using feature extraction weighting method.

- Designing personal Health record-based model for diagnosis of diabetes disease using real data.

1.4 Thesis Organization

The entire thesis work is divided into seven chapters. The brief description regarding these chapters is listed as

- Chapter 1: This chapter describes the introduction of work, motivation and objectives.
- Chapter 2: This chapter present the literature review on different machine learning techniques in the filed disease diagnosis and prediction. The entire literature review divides into two subsections. The first subsection explores the recent works machine learning techniques for various diseases diagnosis and prediction. While, second subsection introduces the recent works reported on diabetes diagnosis and prediction.
- Chapter 3: This chapter presents a KHM-AW-SVM based model for diagnosis and prediction of diabetes diseases. The KHM-AW-SVM model consists of two technique i.e., KHM for determining the optimal attribute for prediction of diabetes, while SVM classifier is used for accurate prediction of diabetes disease through selected attributes through KHM technique.
- Chapter 4: This chapter describes the ABC-DNN based model for diabetes prediction and diagnosis. In ABC-DNN model, ABC is used for selecting optimum features for diabetes disease. Whereas, DNN technique is adopted for prediction of diabetes.
- Chapter 5: The chapter five presents a hybrid diabetes disease prediction framework. The proposed frame work is the combination of K-Mean++ data imputation technique, ABC based outlier method and SVM based classifier.

- Chapter 6: This chapter presents a rule-based monitoring system for diagnosis of diabetes. Further, the real-world diabetes data is collected through an app. Several rules are designed for accurate diagnosis of diabetes disease.
- Chapter 7: This chapter concludes the entire work and provides future direction.

CHAPTER 2

REVIEW OF LIERATURE

People throughout the world are suffered with different types of diseases. Several healthcare systems are developed for effective diagnosis of disease due to development in the field of information and communication technology. These systems aim measure the symptoms of disease and also provide preliminary treatment in case of emergence. In present, different deadliest diseases are occurred, but among these, diabetes is one of common disease responsible for kidney failure, blindness, heart attack etc. The entire literature is divided into two parts

- ML Techniques for Disease Prediction and Diagnosis
- ML Techniques for Diabetes Prediction and Diagnosis

2.1 ML Techniques for Disease Prediction and Diagnosis

This section presents the diverse machine learning techniques for the diagnosis and prediction of various diseases such as heart disease, cancer, stroke, dengue etc.

Lin et al. [22] developed a clinical diagnostic score system for the prediction of coronary artery spasm in patients. In this system, multi variable analysis is performed to identify the patients with coronary disease. The performance of the proposed system is evaluate using nine hundred seventy-six patients. It is noticed that proposed diagnostic score system achieves ninety-six percent accuracy rate.

An ensemble learning framework is designed to diagnosis the thromboembolism [23]. The proposed framework consists of VTE ontology, semantic extraction, sentiment assessment and an ensemble classifier. The ensemble classifier is the combination of the MLPNN and SVM. The performance of proposed framework is investigated on clinical data of two hundred fifty patients. It is seen that proposed ensemble framework obtains higher F-measure rate as compared to other classifiers.

Ahmed and Acharjya adopted cuckoo search and rough set theory to design a hybrid approach for the accurate prediction of heart disease [24]. The cuckoo search technique is used to determine the main features of heart disease. Whereas, rough theory is applied to develop the inference rule for heart disease prediction. The performance of proposed hybrid approach is tested on six hundred three patients. It is observed that proposed approach develops fifty-three rule for the prediction of heart disease and provides state of art prediction results.

Abdar et al. [25] presented a hybrid machine learning technique i.e., N2Genetic optimizer for accurate prediction of coronary artery disease. It is seen that N2Genetic optimiser is used as training technique for SVM technique, called N2Genetic-nuSVM. The well-known Z-Alizadeh Sani dataset is considered to evaluate the performance of the aforesaid technique. Authors claimed that proposed N2Genetic-nuSVM technique achieves 93.08% accuracy rate.

Narayan and Sathiyamoorthy developed a recommender system for prediction and identifying the heart disease [26]. The proposed recommender system is the combination of the Fourier transformation and machine learning technique. In this work, three machine learning technique is used for prediction of heart disease. These techniques are artificial neural network, naïve bayes and support vector machine. The performance of the proposed recommender system is evaluated using real time dataset. The experimental results showed that proposed recommender system is more capable to predict and identify the heart disease as compared to existing ones.

Bucholc et al. [27] designed a computerized decision support system for predicting the accuracy of the Alzheimer's disease. The proposed decision support system consists of six machine learning techniques. These machine learning techniques are Kernel Ridge Regression (KRR), Support Vector Re- gression, and k-Nearest Neighbor for regression and Support Vector Machine (SVM), Random Forest, and k-Nearest Neighbor. It is observed that KKR and SVM provide better classification rate.

An ensemble neuro fuzzy technique is employed for diagnosis of hepatitis disease [28]. The proposed ensemble method works in three steps. In first step, Non-linear Iterative Partial Least Squares method is applied for dimension reduction. In second step, self-organize map method is used for clustering task. In third step, neurro fuzzy inference system is used for predicting the hepatitis disease. Authors claimed that proposed ensemble neuro fuzzy technique provides superior results than Neural Network, ANFIS, K-Nearest Neighbors and Support Vector Machine techniques.

Zhao et al. [29] developed an integrated model based on the logistic regression and support vector machine for the detection of colorectal cancer. In the proposed model, logistic regression method is applied to capture the more relevant features of colorectal cancer. Whereas, SVM with different kernel functions such as linear, sigmoid, radial basis and polynomial are adopted for the detection of colorectal cancer. It is observed that SVM with radial basis function obtains higher accuracy rate as compared to rest of these.

Nilashi et al. [30] designed a medical decision support system for monitoring and diagnosis of real time Parkinson’s disease patients. The proposed medical decision support system is the combination of dimension reduction technique and ensemble learning method. In the proposed system, singular value decomposition technique is implemented as dimension reduction technique. Whereas, adaptive neuro-fuzzy inference system is adopted for the diagnosis of

Parkinson's disease patients. The experimental results showed that proposed system is effectively used to diagnosis and monitoring of Parkinson's disease patients.

Ali et al [31] developed an automated diagnostic system based on the statistical method and deep neural network for prediction of heart disease. The statistical method is used to eliminate the irrelevant features of heart disease dataset whereas, deep neural network is configured for the prediction of heart disease. The performance of automated diagnostic system is compared with ANN and DNN models. It is stated that proposed diagnostic model obtains superior results than ANN and DNN models.

An intelligent risk prediction system based on fuzzy temporal rules is presented for detection of breast cancer [32]. The proposed prediction system employs the fuzzy rule based classification for detection of breast cancer. Some temporal constraints are also imposed in this system. It is observed that proposed risk prediction system computes more accurate risk for breast cancer.

Gokulnath and Shantharajah presented a hybrid technique based on genetic algorithm and support vector machine for the prediction of heart disease, called GA-SVM [33]. In this technique, genetic algorithm is used to select the more relevant features of heart disease. Whereas, SVM is implemented for prediction of heart disease. Cleveland heart disease dataset is used to evaluate the performance of the GA-SVM technique. It is observed that proposed technique provides more accurate results as compared to rest of techniques.

Kirk et al. [34] developed a decision making system called DEAR (Detect, Evaluate, Assess and Recommend action). The basic aim of this proposed system is to detect the change in the environment, to provide the risk assessment and gives real-time recommendations to control the outbreak of mosquito diseases.

Devarajan et al. [35] implemented a healthcare system for the treatment of the Parkinson's disease. The proposed model analyzed the voice samples of the patients to recommend the proper treatment. Fog computing is used as middle layer between the cloud server and end user in the proposed system. Further the Fuzzy K-nearest Neighbor classifier (FKNC) and Case-based Reasoning classifier (CBRC) is utilized to classify the non- Parkinson patients and Parkinson patients. The proposed system also generated an immediate alert in the case of abnormality. The proposed system is tested on the UCI-Parkinson dataset using Accuracy, F-measure, Sensitivity, Specificity and Precision parameter. The experimental results showed that the performance of proposed system is better than existing healthcare systems.

Peker presented a decision support system for improving the accuracy of medical diagnosis [36]. In this work, a new attribute selection method based of k- medoids clustering is applied for attribute weighting. In classification phase, SVM method is applied for diagnosis of medical data. The effectiveness of proposed DSS is tested on three medical datasets i.e. Patkinson's Disease, Heart Disease and Liver Disorder. It is stated that combination of k-medoids and SVM approach outperforms than reported methods in literature.

Huda et al. [37] developed a hybrid ensemble method for imbalance medical data especially for brain tumor diagnosis. The hybrid ensemble method is the combination of feature selection technique and ensemble classification approach. ANN with input gain measurement approximation is applied for selecting better attributes to improve the accuracy of ensemble classification method. In classification stage, decision tree based bootstrap method is applied to classify the imbalance data. Authors claimed that proposed hybrid ensemble method achieves higher accuracy rate in comparison to other popular machine learning algorithms.

Alickovic and Subasi designed a hybrid classification algorithm based on genetic algorithm and rotation forest for breast cancer diagnosis [38]. This algorithm works in two stages. In first stage, genetic algorithm is applied to determine relevant set of features. In second stage, rotation forest algorithm is adopted for diagnosis of breast cancer. Two well-known breast cancer datasets are considered to evaluate the effectiveness of proposed algorithm. It is noticed that proposed algorithm obtains 99% accuracy rate.

Abdar et al. [39] developed a new machine learning technique for an accurate diagnosis of coronary artery disease. In this work, normalization method is applied for pre-processing of data. Further, particle swarm optimization and genetic algorithm are implemented twice for optimizing the results of new optimization technique called N2Genetic optimizer. Z-Alizadeh Sani dataset is used for evaluating the performance of aforementioned algorithm. It is revealed that N2Genetic-nuSVM obtains 93.08% accuracy rate.

Yadav and Jadhav [40] designed an innovative framework for classification of cardiac arrhythmia patients. In the proposed framework, random forest technique is implemented for feature selection as well as prediction. MIMIC-III dataset is considered to evaluate the performance of proposed framework and results are compared with genetic algorithm and grid search techniques. Authors claimed that proposed framework achieves far better results than grid search and genetic techniques in terms of accuracy rate.

Masood et al. [41] developed a computer assisted diagnosis system for detecting the lung cancer. The proposed diagnosis system is the combination of deep learning model and metastasis information. The metastasis is taken from Medical Body Area Network. The performance of proposed system is compared with convolutional neural network. It is stated that proposed diagnosis system obtains 84.88% accuracy rate.

Yadav and Pal [42] applied various ensemble classifiers for accurate diagnosis of thyroid disease. In this work, authors consider bagging, stacking, boosting and voting ensemble classifiers and results of these classifier is evaluated using sensitivity, specificity and accuracy parameters. It is claimed that stacking ensemble classifier provides more accurate results than others.

Liu et al. [43] adopted two methods for predicting the cerebral stroke based on the physiological data. In this work, random forest regression method is used as missing imputation method for computing the missing values of stroke dataset. Furthermore, deep learning technique is applied for the prediction of stroke outcome. Simulation results showed that the deep learning technique achieves minimum false negative rate i.e. 19.1% as compared to other approaches.

Fang et al. [44] developed an integrated machine learning approach to determine the relevant features as well as effective intervention and treatment of Ischemic stroke. In this work, authors consider the international stroke dataset. The relevant features are identified using Shapiro-Wilk algorithm and Pearson correlations. Several machine learning algorithms like VC, MLP, random forest and AdaBoost are used for prediction task. Authors claimed that the feature selection technique improve the accuracy rate Ischemic stroke dataset.

Chen et al. [45] developed a new classifiers based on greedy step wise method and decision tree for improving the accuracy rate. In this classifiers, greedy stepwise method is used to determine relevant attribute, whereas, decision tree technique is applied for prediction task. The performance of proposed classifier is evaluated on Japanese stroke patient dataset. It is revealed that the abovementioned combination obtains more accurate rate than other algorithms.

Govindarajan et al. [46] developed a prototype for classifying the stroke disease based on text mining techniques and machine learning approaches. Authors determine the semantic and syntactic relationship among various attributes of stroke disease. A total five hundred seven patients are considered in this work. The symptoms of stroke disease are taken from the patient's medical sheet. The tagging and entropy approaches are adopted for mine the information from medical sheet. Furthermore, artificial neural network, SVM, boosting, random forest and bagging methods are used for predicting the stroke patients more accurately. Results showed that artificial neural network provides more accurate results as compared to SVM, boosting, bagging and random forest methods.

Arslan et al. [47] applied different data mining techniques for predicting the Ischemic stroke. The dataset contains medical information of eighty stroke patients and one hundred twelve healthy persons, Further, this dataset consists of sixteen features for predicting Ischemic stroke. Three machine learning classifiers such as support vector machine (SVM), stochastic gradient boosting (SGB) and penalized logistic regression (PLR) are used to classify the patients in stroke affected and healthy. Authors claimed that SVM techniques obtains more promising results than other two techniques.

Pas and Goyal [48] explored the applicability of long short term memory with recurrent neural network (LSTM-RNN) for diagnosis of stroke disease. The simulation results are evaluated using accuracy, recall and f-measure parameters. Results confirmed that LSTM-RNN is an effective technique for predicting the stroke outcome.

Li et al. [49] develop a model to predict stroke-associated pneumonia in Chinese AIS patients using machine learning methods. The five machine learning technique are adopted for accurate prediction of stroke. These techniques are logistic regression with regulation, support vector machine, random forest classifier, extreme gradient boosting and fully-connected deep neural network. It is observed that extreme gradient boosting technique provides more promising stroke results as compared to rest of techniques.

Heo et al. [50] applied several machine learning technique to determine the outcome of Ischemic stroke of effective treatment of patients. Authors consider deep neural network, random forest, and logistic regression techniques for the same. The performance of machine learning technique is evaluated using two thousand forty three patients and it is noticed that deep learning technique predicts more accurate outcome of Ischemic stroke as compared to random forest and logistic regression techniques.

Liu et al. [51] developed multi neural network model for prediction and prevention of stroke disease. Authors consider convolutional neural network for determine the relevant features for effective prediction. Further, VGG19, DenseNet, ResNet50 and VGG16 network models are applied for prediction purpose. It is seen that VGG16 model achieves higher accuracy rate than others models.

Monteiro et al. [52] applied machine learning to improve the prediction of functional outcome in ischemic stroke patients. This study showed that machine learning approach achieves marginal accuracy as compared ASTRAL, DRAGON and THRIVE tools. But, it is noticed that adding more features in prediction task improves the performance of machine learning approach significantly. This study focuses on the adding of more features during the treatment of patients.

Shah et al. [53] combined and analyzed the real- time data with historical data of patients to recommend the diagnose. Further, author investigates the challenges in the quality of services for health care applications.

Nandyala et al. [54] designed u-healthcare monitoring system based on IoT and Cloud to Fog(C2F) computing. The proposed system provides more interaction between hospitals and smart homes through end points. It is observed that the proposed system provides quick processing with fewer delay as compared to cloud based systems and can be fulfilled all requirements of emerging models.

Costanzo et al. [55] developed a reliable and flexible health care monitoring system for disease diagnosis. The proposed system is the combination of wearable devices and embedded systems. The aim of proposed system is to monitor the remotely located patients through mobile phones. The proposed monitoring system is also capable to communicate with first aid software's for fast rescue of patients in case of emergency. Moreover, user model based ontology is used to collect the patient data. This data is used by doctors and first aid software's to make the health monitoring system as well as interoperable for users living at home. Further, some fuzzy rules are designed for the diagnostic system. Overall, the aim to proposed monitoring system is to recommend the best treatment in case of critical health conditions.

Oluwagbemi et al. [56] implemented the Ebola fuzzy informatics system using expert systems and fuzzy logic. The proposed system is designed to diagnosis and recommend the Ebola Virus Disease. An online survey also conducted in this regard. It is observed that 31% people have not any information regarding the treatment of Ebola disease. 28% people believed that it can be cured. 43% people agreed that Ebola is a water and air- borne disease while 33% people are disagreed. It is also seen that 24% people have no information regarding transmission of Ebola disease. Whereas, 23% people know that it is transmitted through mosquitoes and insects while, 30% people were ignorant. A comprehensive test is also conducted to assess the performance of proposed system using forty-five people. Moreover, 61% people certified that the proposed system is capable to recommend the treatment of Ebola Virus Disease. Whereas, 16% people are disagreed with the aforementioned statement and rest of have different opinion. Further, 67% people found the proposed system is easy to use while 13% and 20% peoples are disagreed and indifferent.

Sood et al. [57] designed a health care system to monitor and differentiate the various mosquito borne diseases. The core component of the proposed system is fog computing, IoT sensors and cloud computing. The objective of proposed system is to control the diseases at early stage. The proposed system computes the similarity factors to differentiate the diseases. Moreover, the infected users are identified using J48 decision tree classifier. In case of abnormality, fog layer is responsible to generate alert messages and sent to patient mobile. Further, Temporal Network Analysis and Radio Frequency Identification is used to monitor the current state and proximity data. The experimental results showed that the proposed system achieves higher accuracy rate.

Thota et al. [58] designed a security architecture for remote healthcare systems. The proposed architecture provides the asynchronous communication between the data servers and health care applications in the cloud environment. The basic aim of proposed architecture is to track and identify the mobile and other devices in the system, secure the authorization and authentication for all devices.

Venckauskas et al. [59] proposed secure self-authenticable transfer protocol for fog based eHealth architecture. The proposed protocol provided the communications between Fog nodes and Edge nodes. The proposed protocol is utilized as a secure transport for Constrained Application Protocol (CoAP) in the place of Datagram Transport Layer Security (DTLS) and User Datagram Protocol (UDP). The proposed protocol transferred the data flow management and user data authentication information by using the modified header fields of traditional UDP. The experimental result showed that proposed protocol works better in lossy network and with CoAP block transfer mode rather than DTLS and UDP.

Vijayakumar et al. [60] implemented a healthcare system to control the mosquito diseases at early stage. The proposed system used the IoT sensors and wearable to collect the patient information. Further the fog computing is utilized to categorize, share and analyze the medical information between the healthcare service centers and users. It also calculated the similarity coefficients to differentiate the diseases and users are dived into uninfected and infected category using fuzzy k-nearest neighbor approach. Additionally, the Social Network Analysis is employed over the cloud layer to signify the outbreak of diseases. Probability of Disease Outbreak parameter is used to calculate the likelihood of the user. The results showed that proposed system achieves the 95.9% classification accuracy.

Saxena et al. [61] provided an outline of the various treatments on the basis of Ayurveda, complementary and modern homeopathic medicines for ZIKV virus. The author also presented experimental therapeutics for ZIKV infection. Ginier et al. [62] presented that Zika fever may be misdiagnosed as dengue fever, however fever is mild and infrequent in Zika infection. It is noticed that the patient infected with Zika infection just recognizable with skin rashes and minor oedema.

Pabbaraju et al. [63] implemented a reverse transcription polymerase chain reaction (RT-PCR) assay for the detection of Zika, Dengue and Chikungunya virus. The RT-PCR assay is employed for the patient blood testing to recognize and differentiate theses virus for proper treatment. The proposed assay is developer using 3′ untranslated from Dengue, hydrolysis probes targeting the non-structural 5region of Zika and non-structural protein region of

Chikungunya virus. Further the performance of RT-PCR assay is evaluated using standard deviation and coefficient of variation. The results stated the RT-PCR assay is 100% specific and didn't intensify any of the different viruses tested.

Campion et al. [64] designed a web interface to represent the details about occurrence data of west nile virus, mosquito density and weather. The proposed interface used multilayered google maps. The author also developed a prediction model using Partial Least Square Regression technique to predict the mosquito trap counts using data of trap count from 2005-2015 and historical meteorological. The accuracy of proposed model is measured through Mean Absolute Error parameter. The results showed that the proposed prediction models achieve 3.3 statistical accuracy of Mean Absolute Error.

Lambert et al. [65] developed an age monitoring system to predict the accurate age of mosquito. The proposed system deployed various machine learning techniques such as boosted regression trees, principal components regression, random forests and neural networks with Near-Infrared Spectroscopy. Further, the author also suggested the age of mosquito is an important to metric for killing the adult mosquitos. This unprejudiced age estimation creates exact mosquito populace.

Kirk et al. [66] developed a decision-making system called DEAR (Detect, Evaluate, Assess and Recommend action). The basic aim of this proposed system is to detect the change in the environment, to provide the risk assessment and gives real-time recommendations to control the outbreak of mosquito diseases.

Devarajan et al. [67] implemented a healthcare system for the treatment of the Parkinson’s disease. The proposed model analyzed the voice samples of the patients to recommend the proper treatment. Fog computing is used as middle layer between the cloud server and end user in the proposed system. Further the Fuzzy K-nearest Neighbor classifier (FKNC) and Case-based Reasoning classifier (CBRC) is utilized to classify the non- Parkinson patients and Parkinson patients. The proposed system also generated an immediate alert in the case of abnormality. The proposed system is tested on the UCI-Parkinson dataset using Accuracy, F-measure, Sensitivity, Specificity and Precision parameter. The experimental results showed that the performance of proposed system is better than existing healthcare systems.

Kaur et al. [68] proposed a health monitoring system using cloud computing, various machine learning algorithms and IOT infrastructure. The proposed system offered the recommendations for the diagnoses based on the historical data that is lying on the cloud. The proposed system also helped in making the decisions to disguise the various patterns in the database. Further, the author compared the performance of prediction model using accuracy parameter. The different datasets of various diseases and various machine learning algorithms such as Random Forest, MLP, Support Vector Machine, K-NN and Decision Trees are used to compare the performance of prediction model. The name of these datasets are liver disorders, surgical data, breast cancer, heart diseases, spect_heart, diabetes, thyroid and dermatology data. The experimental results showed that prediction model with Random Forest technique achieves 97.26% accuracy on dermatology dataset.

Parthasarathy et al. [69] designed a leg movement monitoring (LMM) system to detect the staring time of illness of joint pain. The proposed LMM system utilized uric acid sensors and wearable sensor gadgets as a part of IoT infrastructure for inflammation ailment of joint. The proposed system is also used to transform the heath data and recognize the movement of human foot for the diagnosis of GOUT arthritis. The performance of proposed system is measured to calculate the time wrapping of Arthritis ailment using specificity and sensitivity parameters.

The results showed that time wrapping technique is dynamic and adequate. The ROC testing is also conducted to differentiate laboratory and clinical assessments to acquire kidney ailment and GOUT arthritis inflammation.

Tuli et al. [70] developed a new model named as HealthFog for automatic analysis of heart diseases. The HealthFog integrated the deep learning (DL) in edge computing (EC) devices. Further the proposed model also provided the services of fog using IoT devices and according to user request it also managed the data of patient. The performance of HealthFog is measured using FogBus in terms of execution time, power consumption, accuracy, latency, network bandwidth and jitter. The results showed that HealthFog provides best prediction accuracy and quality of services.

Priyadarshini et al. [71] a new healthcare model called as DeepFog to predict the status of wellness. The DeepFog is the combination of fog computing and deep learning technique. It collected the patient's data using fog computing and predict the three wellness such as stress type, hypertension attacks and diabetes using deep neural network. The performance of DeepFog is calculated using accuracy, precision, recall and F1 score parameter on standard datasets. The results of DeepFog is compared with other existing systems. The results showed that the DeepFog is efficient for monitoring fitness criteria of three wellness as compared to existing systems.

Jabeen et al. [72] developed an recommender system to diagnoses the cardiac disease. The basic work of proposed system is to recommend the dietary and physical plan to users. The proposed system is divided into four parts. In the first part, the patient's data is collected using bio sensors and data is transferred to the server using IoT environment. In the second part, the features are selected from the above data using a sequential forward selection. In the third part, the machine learning classifiers such as RF, MLP, NB and SVM are utilized to classify the heart diseases into eight cardiovascular classes. In fourth part, the dietary and physical plan is recommended according to the disease age and gender. The proposed system is tested on cardiologist dataset. The performance is evaluated using MBR, recall and precision parameter. The results stated that proposed system achieves 98% average accuracy.

Sood et al. [73] developed a cyber-physical localization (CPL) system based on the cloud computing and neuro fuzzy inference. The basic aim of proposed system is to identify the risk of coronary heart disease, to monitor the electrocardiogram readings of patients, send the alert message to the user mobile and professionals in the case of abnormality in readings and recommend the medication and prevent measure according to risk category. The Adaptive-network-based fuzzy inference system is used to classify the risk level of coronary heart disease. The results stated that the proposed CPL system is efficient in classifying the risk level as well as generates alerts in minimum response time according to electrocardiogram readings.

Gu et al. [74] developed a diagnostic knowledge model (DKM) to classify the clinical criteria, so that deficiency of domain knowledge is reduced between computer scientists and physicians. The basic aim of proposed system is to assuage medical staff from the overwhelming burden of hospital activities and to give opportune decision support. The proposed model utilized the component-based medical cyber-physical system framework (CBMCPS) to incorporate the medical devices and use the knowledge systems. The proposed model is evaluated on 128 clinical cases and compared with existing models. The experimental results stated that proposed model detect the state earlier in 11.02% patients i.e.16.57 hours earlier than existing models.

Sood et al. [75] designed a healthcare system to predict the early stage hypertensive patients from the heath parameters of users. The proposed system continuously analyzes and monitor the blood pleasure patients. The proposed system is divided into four phases. In the first phase, user's data is collected using IoT sensors at fog layer. In the second phase, artificial neural network is used to predict the risk of hypertension attack. In the third phase, the alerts are generated in the case of abnormality in blood pressure and sent to mobile. In the fourth phase, the medical information of patients is stored on cloud for doctors, personal caregivers and clinicians for recommending the precaution measures at early stages. The results demonstrated that proposed system achieves bandwidth efficiency, low response time and high accuracy.

Lakshmanaprabu et al. [76] developed a healthcare system using MapReduce, IoT infrastructure, Improved Dragonfly Algorithm and RF classifier. The proposed system is divided into two phases. In the first phase, patient's data is collected using IoT sensors and MapReduce process. In the second phase, improved Dragonfly Algorithm is used to select the attributes from the dataset. In the third phase, RF classifier is utilized to classify the different diseases using selected attributes. The proposed system is evaluated on different real time hospital datasets using precision parameter. The performance of proposed system is compared with existing systems and achieves 94.2% precision.

Anand et al. [77] designed a hybrid model to classify the liver syndrome. In the first stage, the medical data is classified on the basis of existence diseases. In the second stage, the modified particle swarm optimization algorithm is developed to sect the features from the medical dataset. In the third stage, the modified artificial neural network is employed to classify the diseases. The performance of proposed system is evaluated using Spark tool and results demonstrated that proposed system improves the classification accuracy as compared to existing systems.

Devarajan et al. [78] implemented a healthcare system for the treatment of the Parkinson's disease. The proposed model analysed the voice samples of the patients to recommend the proper treatment. Fog computing is used as middle layer between the cloud server and end user in the proposed system. Further the Fuzzy K-nearest Neighbor classifier (FKNC) and Case-based Reasoning classifier (CBRC) is utilized to classify the non- Parkinson patients and Parkinson patients. The proposed system also generated an immediate alert in the case of abnormality. The proposed system is tested on the UCI-Parkinson dataset using Accuracy, F-measure, Sensitivity, Specificity and Precision parameter. The experimental results showed that the performance of proposed system is better than existing healthcare systems.

Kaur et al. [79] proposed a health monitoring system using cloud computing, various machine learning algorithms and IOT infrastructure. The proposed system offered the recommendations for the diagnoses based on the historical data that is lying on the cloud. The proposed system also helped in making the decisions to disguise the various patterns in the database. Further, the author compared the performance of prediction model using accuracy parameter. The different datasets of various diseases and various machine learning algorithms such as Random Forest, MLP, Support Vector Machine, K-NN and Decision Trees are used to compare the performance of prediction model. The name of these datasets are liver disorders, surgical data, breast cancer, heart diseases, spect_heart, diabetes, thyroid and dermatology data. The experimental results showed that prediction model with Random Forest technique achieves 97.26% accuracy on dermatology dataset.

Jabeen et al. [80] developed a recommender system to diagnoses the cardiac disease. The basic work of proposed system is to recommend the dietary and physical plan to users. The proposed system is divided into four parts. In the first part, the patient's data is collected using bio sensors and data is transferred to the server using IoT environment. In the second part, the features are selected from the above data using a sequential forward selection. In the third part, the machine learning classifiers such as RF, MLP, NB and SVM are utilized to classify the heart diseases into eight cardiovascular classes. In fourth part, the dietary and physical plan is recommended according to the disease age and gender. The proposed system is tested on cardiologist dataset. The performance is evaluated using MBR, recall and precision parameter. The results stated that proposed system achieves 98% average accuracy.

Tuli et al. [81] developed a new model named as HealthFog for automatic analysis of heart diseases. The HealthFog integrated the deep learning (DL) in edge computing (EC) devices. Further the proposed model also provided the services of fog using IoT devices and according to user request it also managed the data of patient. The performance of HealthFog is measured using FogBus in terms of execution time, power consumption, accuracy, latency, network bandwidth and jitter. The results showed that HealthFog provides best prediction accuracy and quality of services.

Priyadarshini et al. [82] developed a new healthcare model called as DeepFog to predict the status of wellness. The DeepFog is the combination of fog computing and deep learning technique. It collected the patient's data using fog computing and predict the three wellness such as stress type, hypertension attacks and diabetes using deep neural network. The performance of DeepFog is calculated using accuracy, precision, recall and F1 score parameter on standard datasets. The results of DeepFog are compared with other existing systems. The results showed that the DeepFog is efficient for monitoring fitness criteria of three wellness as compared to existing systems.

2.2 ML Techniques for Diabetes Prediction and Diagnosis

This section presents the related works reported in the direction of diabetes disease prediction.

Yuvaraj and SriPreethaa [83] considered the above-mentioned disease and developed a healthcare system for diabetes prediction. In this work, authors consider a distributed computing framework, called Hadoop for predicting the diabetes. This work implements the machine learning algorithm in Hadoop environment and these algorithms determine the Hadoop cluster for determining the diabetes. This work selects the three popular ML algorithms such as DT, NB and RT and simulation results are evaluated using accuracy, precision and recall parameters. The well-known pima indian diabetes dataset is chosen for assessing the simulation results of distributed framework. The results showed that Hadoop+RT framework obtains more accurate and stable results than Hadoop+NB and Hadoop+DT frameworks. The Hadoop+RT framework achieves an average of 4% higher accurate results than other frameworks.

Xie and Wang [84] presented the comparative analysis of several ML techniques with classical auto regression technique for monitoring the blood glucose. In this work, an autorepression model is reported with exogenous inputs, called ARX model. The ML techniques are characterized using regression, deep learning with LSTM and temporal convolution network (TCN). The experiment is implemented using different input attribute, regression model, recursive model and direct method for multistep diabetes prediction. The OhioT1DM dataset is used for implementing the proposed autoregression model. The simulation results are evaluating using RMSE and temporal gain for early diabetes prediction. While, normalized energy is adopted for predicting the risk of false rate either hypo or hyper glycemia. The simulation results showed that ARX model achieves least RMSE values for both direct and recursive methods. The direct method achieves second highest temporal gain. However, ML algorithms cannot exhibit significant results as compared to autoregression technique.

Hypertension and high blood pressure are life threaten complication of diabetes disease. These symptoms can be undiagnosed and untreated during the diabetes prediction and put severe impact on patients. The aforementioned risks can be overawed through early identification of diabetes.

Zhu et al. [85] developed an improved logistic regression model for diabetes prediction using principal component analysis (PCA) and K-means clustering algorithm. In proposed, PCA is adopted to map the diabetes data in lower dimension. Simulation results showed that integration of PCA is improved the accuracy results of K-means clustering and logistic regression. Moreover, authors claimed that aforementioned methods are successfully adopted for predicting the diabetes patients using Patient Electronic Health Record Data.

Devi et al. [86] integrated the farthest first clustering algorithm and sequential minimization optimization (SMO) classifier for improving the diagnostic results of diabetes mellitus. In this work, farthest first algorithm is used to group the data in to clusters. In turn, computation time is reduced due to shrinkage of data. Further, SMO classifier is applied on the output of farthest first clustering algorithm. This algorithm classifies the data into tested positive and negative. It is observed that integration of farthest first clustering algorithm and sequential minimization optimization improve the accuracy rate of diabetes mellitus.

Maniruzzaman et al. [87] designed a machine learning based system for predicting diabetes patients. In the proposed model, logistic regression is adopted to determine the risk factor associated with diabetes disease based on p value and odd ratio. Further, four classifiers such as naive Bayes (NB), decision tree (DT), Adaboost (AB), and random forest (RF) are used for prediction task. Moreover, three types of partition protocol i.e. K2, K5 and K10 are also used in this study. The performance of classifiers is evaluated using accuracy rate. It is stated that RF based classifier with K10 protocol obtains ninety-four percent accuracy rate as compared to other classifiers.

Wang and Chen [88] developed a boosted support vector machine for improving the classification accuracy of diabetes disease based on chaotic multi-swarm whale optimization algorithm. In this work, multi swarms with stratified mechanism and self-adaptive chaotic disturbance mechanism are integrated in whale optimization algorithm (WOA). Moreover, the proposed whale optimization algorithm is used for parameter optimization of SVM and feature selection. Finally, WOA optimized SVM is applied for the prediction of diabetes disease. Authors claimed that integration of WOA and SVM achieves higher classification results as compared to same class of algorithms.

To detect the diabetes mellitus in early stage, Singh and Singh [89] presented stacking-based evolutionary ensemble learning system i.e., NSGA-II-Stacking for predicting Type-II diabetes mellitus. In the proposed learning system, twenty learning models are created as base models. Moreover, k-nearest neighbor model is applied to combine the prediction outcome of twenty base learner models. Simulation results of proposed learning system are taken over the PIMA indian diabetes dataset. Authors claimed that proposed NSGA-II stacking model outperforms than other individual machine learning algorithms.

Siva and Manikandan [90] developed diabetes prediction model based on grey wolf optimization (GWO) and fuzzy logic. In proposed prediction model, fuzzy logic is applied to generate the fuzzy rules. In this study, seventeen fuzzy rules are generated using eight features and two classes. Further, GWO algorithm is adopted to optimize the generated fuzzy rules. Simulation results showed that GWO algorithm provides better results than ant colony optimization algorithm.

For earlier detection of Gestational diabetes mellitus (GDM), Artzi et al. [91] applied gradient boosting model for GDM prediction. Simulation results are evaluated using area under curve (AUC) parameter. It is stated that proposed gradient boosting model achieves 0.85 percent AUC rate.

Bernardini et al. [92] presented sparse balanced support vector machine (SB-SVM) for diagnosis of type 2 diabetes mellitus. The effectiveness of the proposed approach is tested over HER dataset. Simulation results are compared with deep learning and machine learning methods. It is observed that SB-SVM outperforms than deep leaning and machine learning methods.

For effective management of diabetes disease, Li et al. [93] developed deep learning model for forecasting of blood glucose level. The reliability of proposed learning model is tested on two types of datasets i.e. simulator and clinical datasets. Authors claimed that proposed deep learning model achieves lower root mean square error as compared to existing models.

Aliberti et al. [94] designed multi-level data driven approach for measuring blood glucose level to monitor diabetes disease. In this work, authors present two different kind of approaches i.e. non-linear autoregressive (NAR) neural network and on long short-term memory (LSTM) networks. Simulation results are compared with feed-forward neural networks (FNNs), autoregressive (AR) models, and recurrent neural networks (RNN). It is seen that NAR approach provides efficient results for short term prediction. While, LSTM provides better accuracy results for both short-term and long-term prediction.

To improve healthcare quality and avoid adverse situation of individuals, Fitriyani et al. [96] designed disease prediction model for early detection of type 2 diabetes. The proposed model is the combination of isolation forest (iForest) based outlier detection, synthetic minority oversampling technique tomek link (SMOTETomek) and ensemble approach. iForest method is used to remove the outlier detection. While, SMOTETomek is adopted for balanced data distribution. Ensemble approach is considered for the prediction task. In this work, multilayer perceptron (MLP), support vector machines (SVM), and decision tree (DT) are acted as first level classifiers. While, logistic regression classifier works as second layer classifiers. Simulation results showed that proposed model obtains more accurate results than other compared models.

Hasan et al. [96] presented a novel framework for diabetes prediction. The proposed framework integrates different tasks for diabetes prediction such as outlier detection, missing values, data standardization, feature selection, K-fold cross-validation, and different machine learning (ML) classifiers. Moreover, weighted ensemble method of different machine learning models is also developed for improving prediction of diabetes. Authors claimed that proposed frame work obtains better results as compared to existing studies.

Chen et al. [97] designed ontology-based system for diabetic patients using detailed studies on diabetes disease for its diagnosis and treatment. In this study, semantic web ontology is used for developing ontology-based diabetes framework. Reliability of proposed system is tested on seven hundred sixty-six medical records. Simulation results illustrated that proposed system achieves more accurate results as compared to existing system.

Bassest et al. [98] developed a framework based on computer propped diagnosis and IoT for detecting and observing of type-2 diabetes. In this work, authors developed a decision-making model for type -2 neutrosophic numbers and VIKOR method. It is stated that proposed recommendation system achieves better accuracy rate.

To improve the prediction rate of diabetes disease, Alirezaei et al. [99] presented a hybrid optimization method for reducing noise and data dimension for diagnosis of diabetes. In their work, K-means based algorithm is applied for detecting and deletion of outlier. Moreover, four bi-objective algorithms are implemented to identify relevant features. Further, SVM classifier is adopted for the prediction task. It is noticed that multi-objective firefly algorithm obtains higher accuracy rate among all other algorithms.

Anuradha et al. [100] designed an effective and efficient a fuzzy rule miner using ant colony optimization for diabetes prediction and diagnosis. The efficacy of rule miner is tested using PIMA indian disease dataset. It is noticed that proposed rule miner obtains 87.7 % accuracy rate.

To address the prediction accuracy of issue, Jayashree and Kumar [101] presented evolutionary correlated gravitational search algorithm (ECGS) to determine the appropriate features for diabetes disease. Further, genetic optimized Hopfield neural network (GHNN) method is used to process the selected features. PIMA indian diabetes dataset is used to evaluate the performance of aforementioned method. It is observed that proposed GHNN method achieves minimum root mean square error as compared to rest of methods being compared.

Kannadasan et al. [102] build deep neural network framework for type-2 diabetes disease classification or prediction using stacked autoencoders. The task of stacked autoencoders is to select the optimum features for prediction of diabetes disease. Simulation results showed that proposed framework achieves 86.26 % accuracy rate.

A case-based preparation framework for CBR systems that converts electronic health record data into fuzzy CBR knowledge is presented in [103]. The proposed model generates fuzzy case base knowledge using standard crisp entity–relationship data model. Simulation results demonstrate that proposed fuzzy CBR is better than traditional CBR system.

Cho et al. [104] considered risk-based clustering method for the identification of type-2 diabetes mellitus. In this work, five risk such as age, gender, body mass index, hypertension, and family history of diabetes are considered from population. The proposed approach identifies six population clusters with different prevalence of type-2 diabetes mellitus.

CHAPTER 3

KHM-AW -SVM DAIGNOSTIC SYSTEM FOR DIABETES

The diagnostic models contain different ML techniques for improving the prediction rate [4-7]. The features of datasets also having significant role in decision making process. The prediction accuracy is also depending on nature of features either relevant or irrelevant. The issue of relevancy and irrelevancy features especially for medical datasets, can be focused by several researchers. In turn, irrelevant features can be removed from the medical datasets and this process is known as feature reduction. It can be defined as to determine the irrelevant features from datasets, because irrelevant features can degrade the performance of classifiers. On the other side, SVM is popular classifier that can widely adopted for solving medical data prediction task [29]. But several drawbacks related with SVM classifier. These are summarized as (1) choose a suitable kernel function, (2) selection of optimal input to SVM classifier, (3) parameters (weight and bias) optimization of SVM. In above-mentioned drawbacks, second drawback is most challenging because it can affect the performances of kernel function and SVM parameters.

3.1 Chapter Contribution

This chapter considers the selection of optimal input issue of SVM classifier and resolves this issue using k-harmonic mean-based attribute weighting method (KHM-AW). The contributions of this chapter are mentioned below.

- Develop a KHM-AW -SVM diagnostic model for improving the diabetes accuracy.
- Design KHM-AW technique for extracting relevant features for SVM classifier.
- Adopt SVM classifier for classification and prediction of diabetes dataset.

3.2 Proposed KHM-AW -SVM Diagnostic Model

This section presents the KHM-AW -SVM diagnostic model. The proposed model comprises of four steps of i) pre-processing step, ii) 10 cross fold, iii) Model construction, and iv) Model usage. The preprocessing step implements the KHM-AW as attribute weighting technique for extracting the relevant features. The subsection 3.2.1 discuss the proposed KHM-AW technique in detail. The second step consists of 10 cross fold method. This method divides the entire diabetes data set in ten equal subsets in which nine subsets are used for train the model and rest subset used as test set. Further, every time test set is different. Third step corresponds to model construction. The model construction can be described as to train/construct model using classifiers. In this work, SVM classifier is adopted for model construction. The detailed description of SVM classifier is discussed in subsection 3.2.2. Fourth step corresponds to model usage. This step evaluates the performance of classifier with respect to test case. The schematic working of proposed KHM-AW-SVM diagnostic model is illustrated in Figure 3.1.

3.2.1 KHM-AW Method

This section explains the details of KHM-AW method. KHM-AW aims to determine the non-relevant attribute in the dataset, because non-relevant features can reduce the diagnostic accuracy of algorithms. In literature, it is mentioned that KHM provides superior results over K-Means and EM algorithms [105]. KHM is also less sensitive toward initial cluster centers and considered harmonic mean to update the cluster centers, in turn obtains improved results in comparison to similar algorithms. Due to afore-mentioned advantages, this work investigates the capability of KHM to discover non relevant attributes for diabetes diseases. The procedure of KHM-AW method can be described as

1. Compute the optimum cluster centers using KHM method.

2. Data are arranged into respective clusters using minimized value of Euclidean distance.

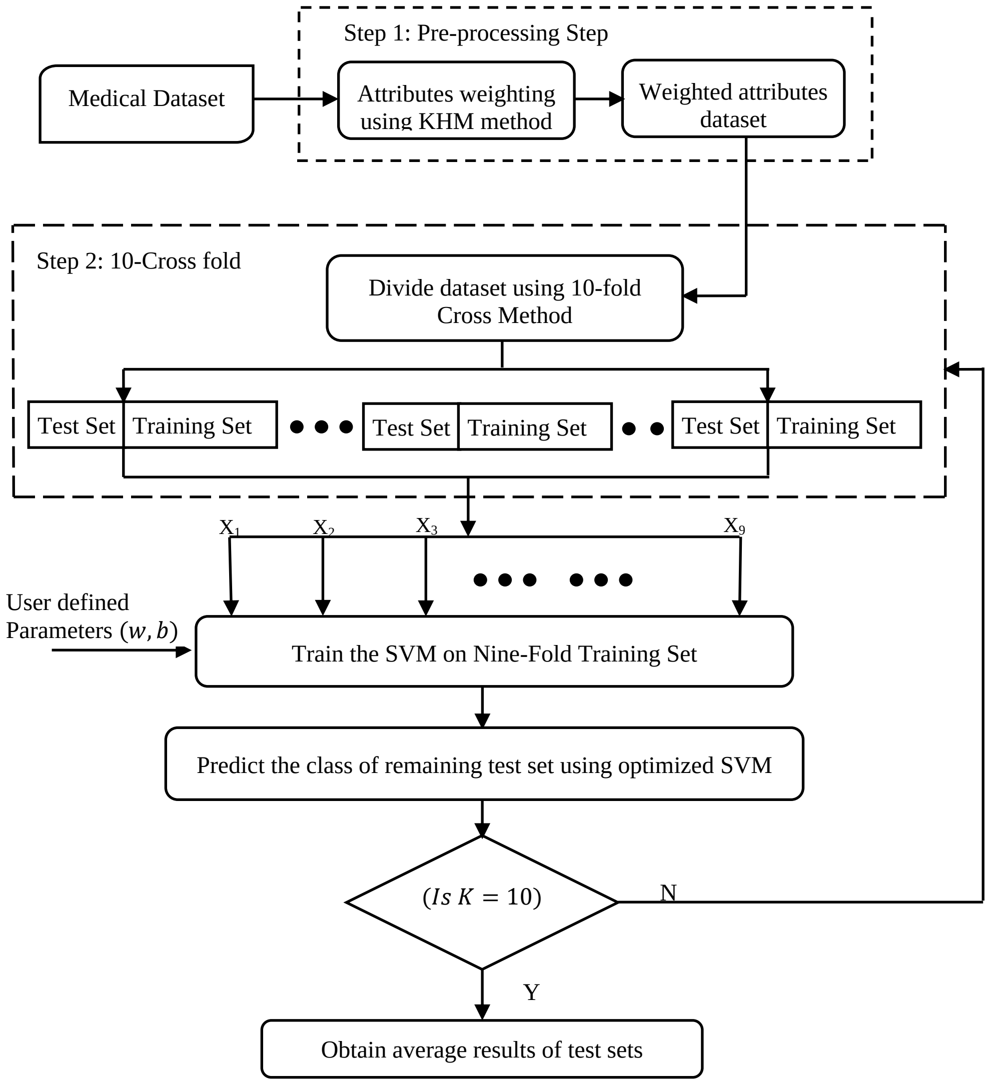


Figure 3.1: Proposed KHM-AW-SVM Diagnostic System

1. Determine the average value of attributes through clusters. It can be measured through two multiplicative functions i.e. F_1 and F_2, mentioned in equations 3.1-3.2.

$$F_1 = \frac{(\sum_{i=1}^{N} x_i)/N}{(\sum_{k=1}^{K} c_k)/K} \tag{3.1}$$

$$F_2 = \frac{(\sum_{k=1}^{K} c_k)/K}{(\sum_{i=1}^{N} c_i)/N} \tag{3.2}$$

2. If, data objects consist of larger value than optimum cluster, then apply function (F_1)
3. Otherwise, apply (F_2).
4. If, computed value similar/near to average value of optimum clusters, then computed value multiply by 1.

Figure 3.2 demonstrates the working behavior of KHM-AW method. Consider, a sample dataset with data objects (n), attributes (m), clusters (K) and attributes are described as $A_1, A_2, A_3, \dots, A_m$. Step 1 corresponds to attributes and its respective values in the sample dataset. Step 2 corresponds for labelling of data objects. So, to determine the labelling of data objects, KHM clustering

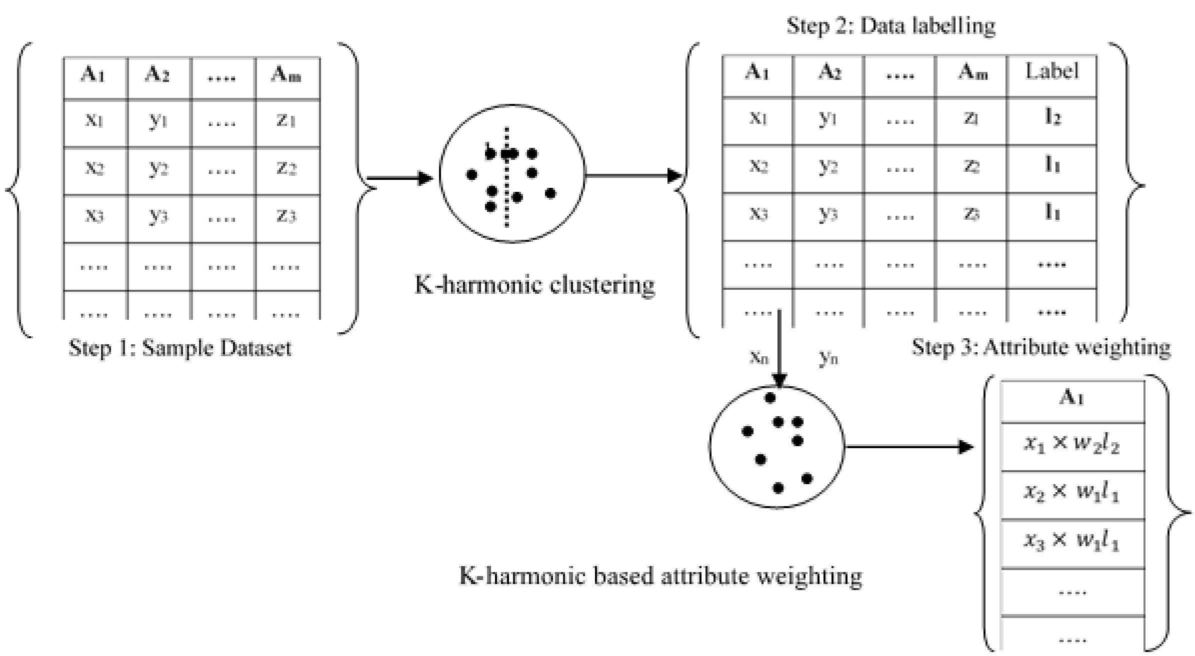


Figure 3.2: Working procedure of KHM-AW method

algorithm is applied and assigned either label (l_1) or label (l_2) to data objects. Moreover, this algorithm also computes two coefficients (w_1) and (w_2). These coefficients multiply with attributes values to compute final weight. Step 3 corresponds for computing the weight of attributes through coefficients (w_1) and (w_2). w_1 is determined using equation 3.3, whereas, w_2 is computed using equation 3.4.

$$w_1 = \frac{F_1}{C_1} \text{ i.e. } F_1 = \frac{\sum_{i=1}^{j} x_i}{j} \tag{3.3}$$

$$w_2 = \frac{F_2}{C_2} \text{ i.e. } F_2 = \frac{\sum_{n=1}^{m} x_n}{m} \tag{3.4}$$

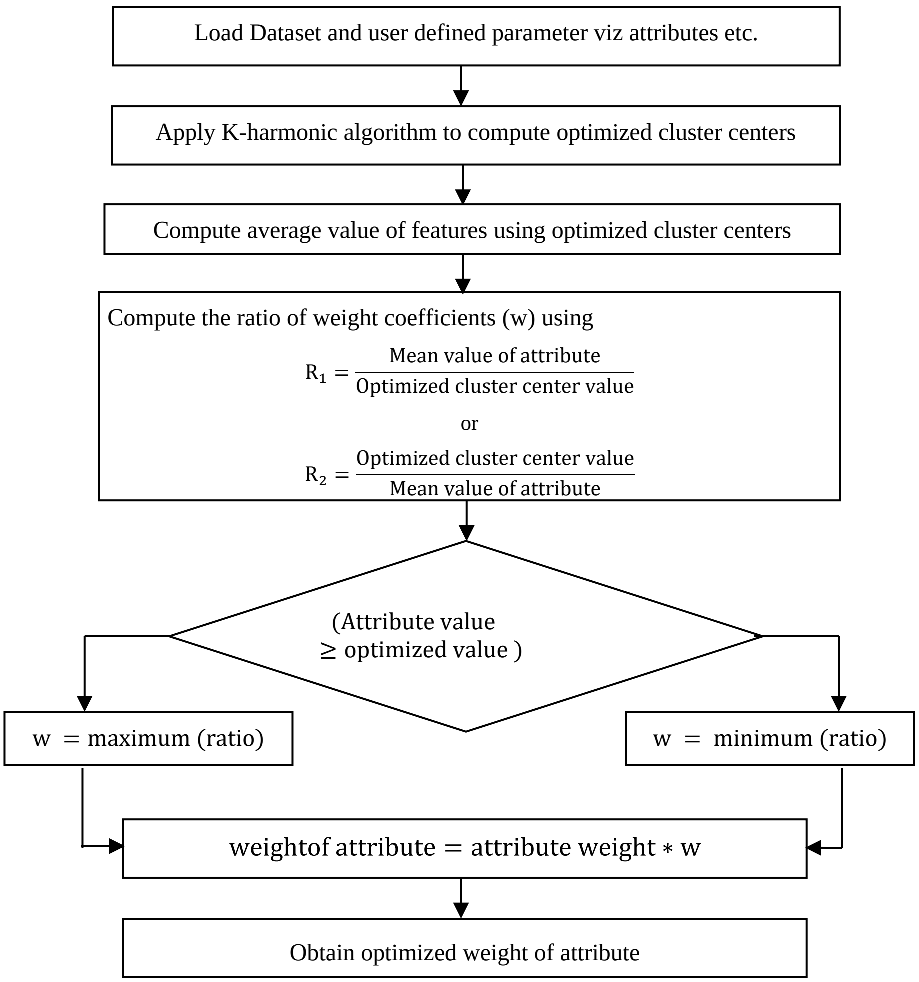


Figure 3.3: Flowchart of KHM-AW method

In equations 3.3-3.4, w_1 and w_2 denote weight coefficient of the attribute A_1, F_1 and F_2 describe average weight of attribute, C_1 and C_2 denote corresponding cluster centers obtained using KHM-

AW method, j and m are number of data objects in l_1 and l_2 respectively. Figure 3.3 presents the flowchart of KHM-AW method.

3.2.2 Support Vector Machine

SVM is a popular classifier than can be used for classification and regression tasks and inspired through statistical learning [29]. This classifier designed for solving two class classification problem and worked efficient with linear data. This classifier divides the data into two classes using an optimal decision function. The optimal function can be interpreted as designing a hyperplane including decision function to separate the data into two classes and $\{-1, +1\}$ represents these classes [29, 33]. The main reason for designing the hyperplane is to plot the data into hyper plane so that distance between data points can be maximized. A two-class problem with n data objects is described as $\{T_i, P_i\}, i = 1,2,3,4, \dots\dots, n$ and inequalities of hyperplane is mentioned in equation 3.5.

$$T_i = +1, wP_i + b \geq +1 \text{ or } T_i = -1, wP_i + b \leq -1 \tag{3.5}$$

The equation 3.5 can be described as P_i is i^{th} data object, T_i is class label of i^{th} data object that can be either +1 or -1, w and b is weight vector and bias. Here, w and b represent the user defined parameters and these parameters also affect the performance of SVM. For improving the SVM performance, it is suggested that plot more than one hyperplane in parallel order as described in Figure 3.4(a) and determine the optimal hyper plane with maximum difference as defined in Figure 3.4(b). The parallel hyper plane is constructed through points, called support vector and it is computed using equation 3.6.

$$wP_i + b = \pm 1 \tag{3.6}$$

Further, w is minimum for achieving the boundary of optimum hyper plane. It also imposes a constraint on hyper plane for effectively computing optimum hyper plane and it is described in equation 3.7.

$$\text{minimum}\left[\frac{1}{2}w^2\right] \tag{3.7}$$

Now, the two-class problem with constraint can be expressed through equation 3.8

$$T_i(wP_i + b) \geq +1 \text{ and } T_i \in \pm 1 \tag{3.8}$$

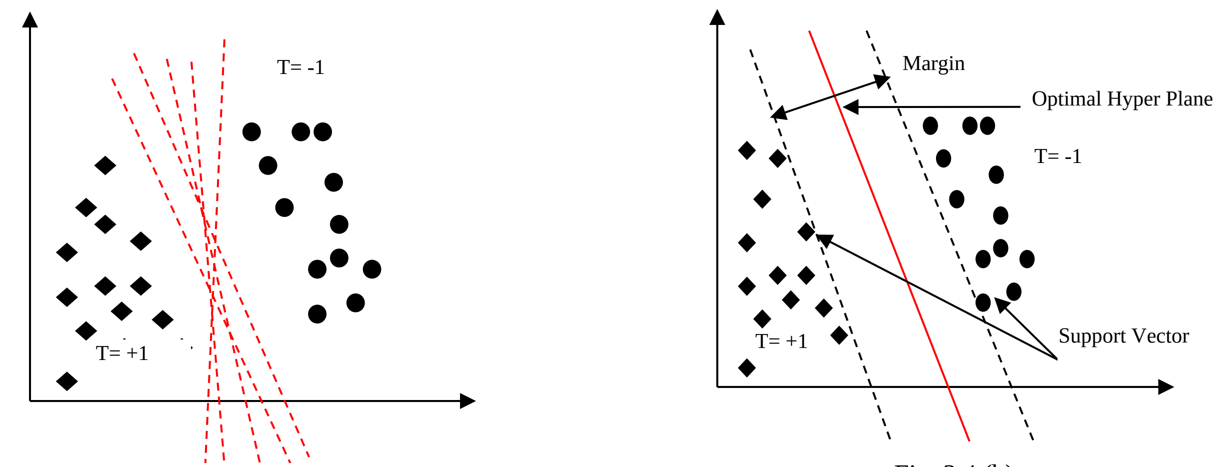


Fig. 3.4 (a) Fig. 3.4 (b)

Figure 3.4(a-b): Illustrate the procedure of SVM using linear data

Further, Lagrange equation can be used for solving constraint two class problem and it can be described through equation 3.9.

$$L(w, b, \beta) = \frac{1}{2}\|w^2\| - \sum_{i=1}^{n} \beta_i P_i (wP_i + b) + \sum_{i=1}^{n} \beta_i \tag{3.9}$$

The decision function for computing the label of class is expressed in equation 3.10.

$$f(P) = \text{sign}\left(\sum_{i=1}^{n} \gamma T_i\,(PP_i) + b\right) \tag{3.10}$$

Sometime, data is non-linear in nature for many problems as described in Figure 3.5(a and b). A new parameter (δ_i) is defined in SVM for solving non-linear data problems and overfitting problem of SVM is also addressed especially in case of hyper plane. Further, for effective balancing among boundary limit and classification errors, a parameter (C) is integrated into SVM and the range of C in between [0, ∞].   So, the SVM for non-linear optimization problems can be formulated using equation 3.11 and after imposing constraint, it is described through equation 3.12.

$$\text{minimum}\left(\frac{1}{2}\|w^2\| + C * \sum_{i=1}^{j} \delta_i\right) \tag{3.11}$$

$$T_i(w * \beta(P_i) + b) - 1 \geq 1 - \delta_i;\ \delta_i \geq 0 \text{ and } i = 1,2,3, \ldots\ldots n \tag{3.12}$$

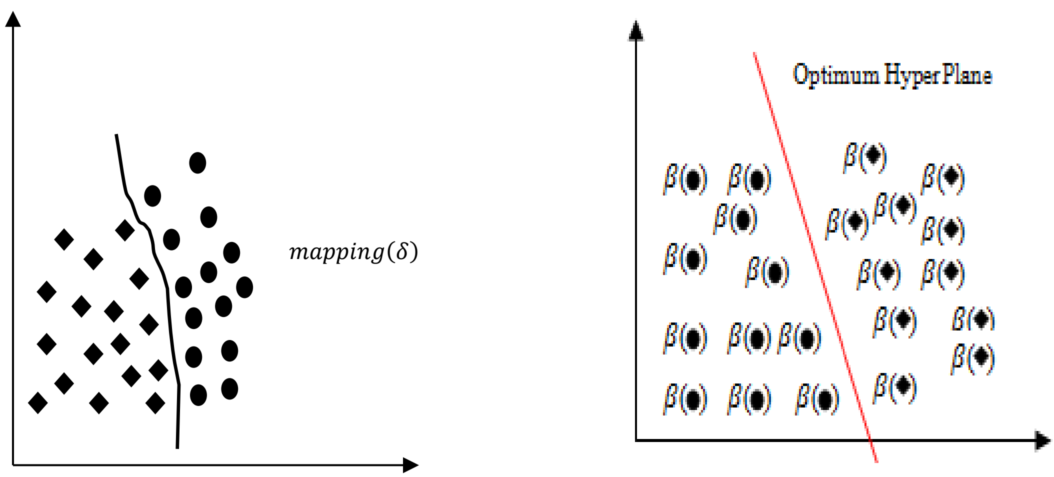


Figure 3.5 (a)   Figure 3.5(b)

Figure 3.5(a and b): Graphical illustration of SVM for non- linearly separable data

It is suggested that non-linear transformation can be handled through kernel function. It can be defined using $K(P_i, P_j) = \beta(P_i) * \beta(P_j)$ and confirmed separation of non-linear data into linear one through high dimensional data space. Moreover, a new decision function is designed for handling the non-linear data and it is described in equation 3.13.

$$f(P) = \text{sign}\left(\sum_{i=1}^{n} \gamma T_i\, \beta(P_i) * \beta(P_j) + b\right) \tag{3.13}$$

The selection of kernel function also affects performance SVM technique, RBF and Gaussian are popular kernel functions. This work considers the RBF as kernel function.

3.3 Experimental Results

The simulation results of KHM-AW-SVM diagnostic model are discussed in this section. The pima Indian diabetes dataset is considered for evaluating the performance of proposed model. The results are validated using 10-cross fold and 50-50% training and testing method. The average of ten run is used for describing the simulation results. SVM parameters are tuned according the existing study [29, 33]. The efficacy of KHM-AW-SVM diagnostic model is measured through diverse performance parameters like accuracy, sensitivity, specificity, AUC, kappa and f-measure. The equations 3.14-3.17 describes the aforementioned parameters.

$$\mathrm{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \times 100 \tag{3.14}$$

$$\mathrm{Sensitivity} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \times 100 \tag{3.15}$$

$$\mathrm{Specificity} = \frac{\mathrm{TN}}{\mathrm{FP} + \mathrm{TN}} \times 100 \tag{3.16}$$

$$\mathrm{F - Measure} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{3.17}$$

In pre-processing phase of KHM-AW-SVM, KHM-AW technique is applied to determine the labelling of data objects. Figure 3.6 illustrates the labelling of data objects into two labels (diabetes and non-diabetes) using KM-AW method. The optimal attributes of diabetes dataset are used for plotting the data distribution. The simulation results of KHM-AW-SVM model is depicted into Table 3.1 using 10-fold and 50-50% training-testing methods. Further, results are evaluated using all attributes and optimum attributes (computed through KHM-AW method) of diabetes dataset. The KHM-AW-SVM model achieves 91.38% accuracy with optimum attributes using 10-fold method, while using all features, the accuracy rate is 81.6%. The KHM-AW method improves the accuracy rate of diabetes disease more than 10%. The proposed model achieves 88.76% accuracy with optimum attributes using 50-50% training-testing method, while using all features, the accuracy rate is 79.24. The simulation results are enhanced the accuracy rate more than nine

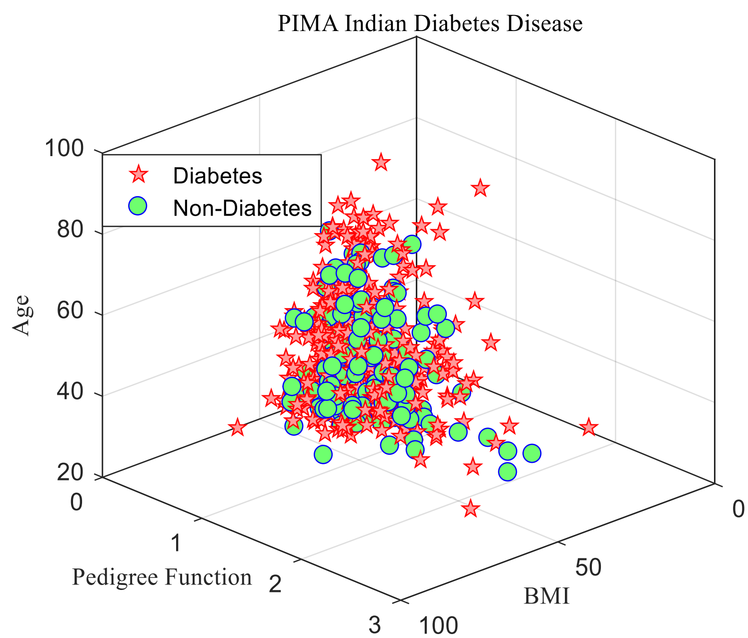


Figure 3.6: Demonstrates the distribution of data into labels using KHM-AW method

percent. It is concluded that incorporation of KHM-AW method into SVM classifier improves the performance in significant manner. On the analysis of other parameters, it conferred that KHM-AW-SVM model achieves more promising results using 10-fold and 50-50% training-testing methods as compared to without attribute selection mechanism. It is also analyzed that there is significant performance difference between the 10-fold method and 50-50% training-testing method. Results conferred that 10-cross method achieves higher accurate results than 50-50% training-testing method.

Table 3.1: Simulation results of SVM and KHM-AW-SVM on diabetes dataset.

Features	Parameters	10 cross fold	50-50% training and test
Using All Features	ACC	81.6 ± 5.26	79.24 ± 6.31
	Sensitivity	87.32± 6.58	84.91 ± 6.46
	Specificity	73.46 ± 6.45	70.36 ± 6.78
	f-measure	0.76	0.72
	Kappa	0.68	0.66
	AUC	0.8	0.78
Using Attribute Weighting	ACC	92.38 ± 4.31	88.76 ± 4.12
	Sensitivity	92.21 ± 5.09	91.54 ± 5.38
	Specificity	90.52 ± 4.93	87.61 ± 4.88
	f-measure	0.85	0.81
	Kappa	0.88	0.83
	AUC	0.89	0.86

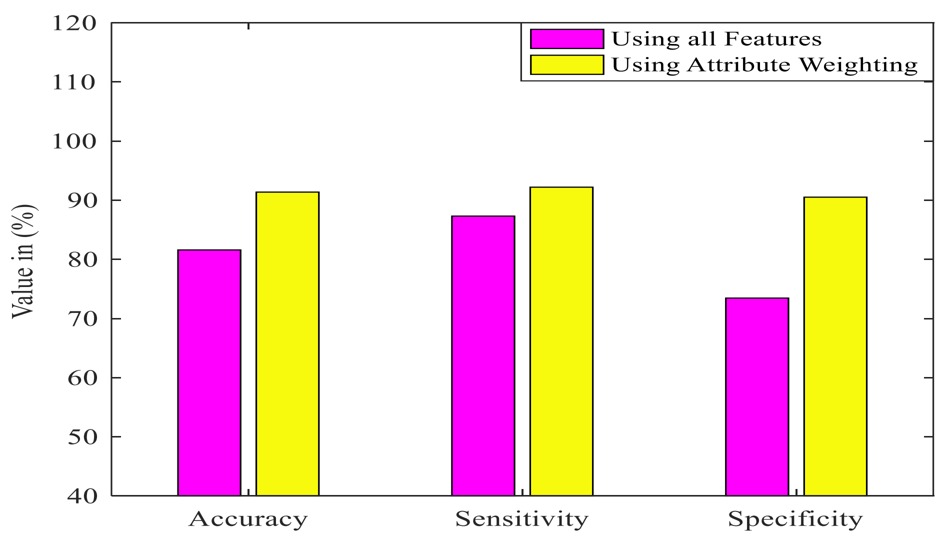
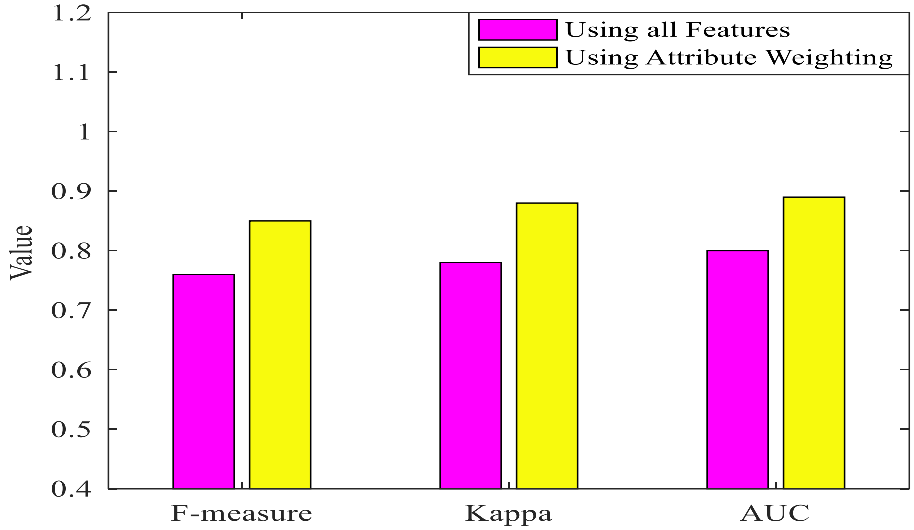
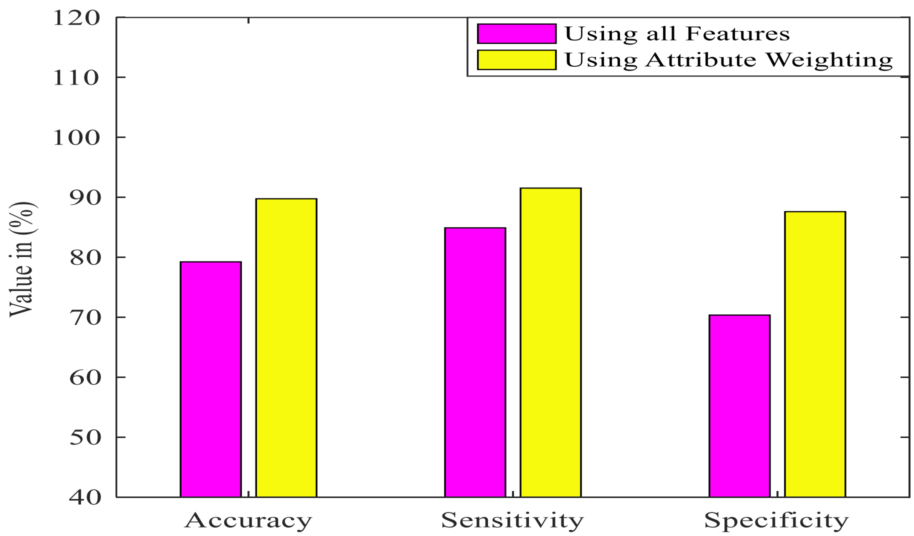


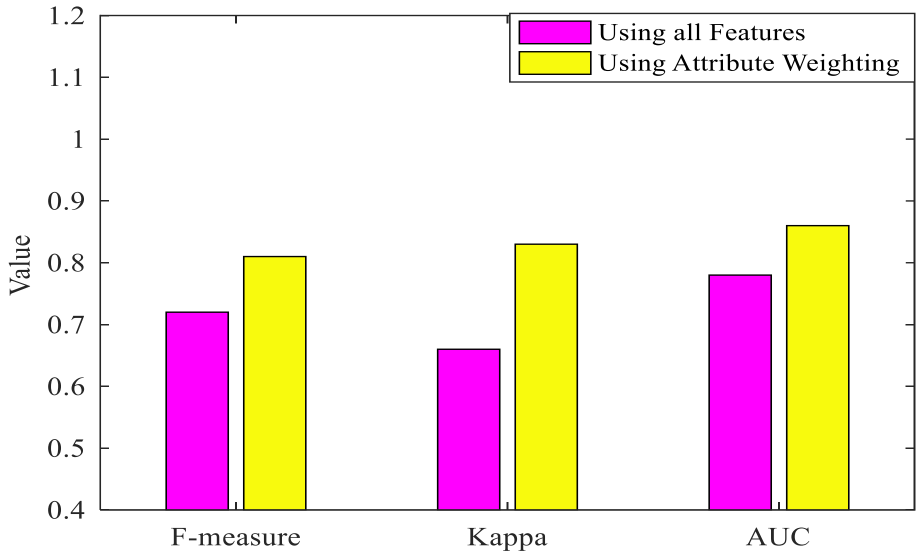
Figure 3.7 Simulation results of the SVM and proposed KhmAW-SVM with 10-fold technique using accuracy, sensitivity and specificity parameters.



Figures 3.8: Simulation results of the SVM and proposed KhmAW-SVM with 10-fold technique for f-measure, kappa and AUC parameters.



Figures 3.9: Simulation results of the SVM and proposed KhmAW-SVM with 50-50% training-testing technique using accuracy, sensitivity and specificity parameters.



Figures 3.10: Simulation results of the SVM and proposed KhmAW-SVM with 50-50% training-testing technique using f-measure, kappa and AUC parameters.

The simulation results of KHM-AW-SVM and SVM is depicted into Figures 3.7-3.8 using 10-fold method. Its stated that integration of KHM-AW method into SVM improves the results of SVM in significant manner. The results of 50-50% training-testing method is reported in Figures 3.8-3.9. It is analyzed that KHM-AW-SVM provides more accurate results than SVM. gives improved results than SVM. Finally, it conferred that attribute weighting method having significant impact on the performance of classifiers.

Furthermore, the results of proposed KHM-AW-SVM model are also compared with existing literature on diabetes disease. As a result, thirty-one studies are chosen from the literature for comparing the results of KHM-AW-SVM model. The accuracy parameter can be considered for comparing the results of all these techniques including KHM-AW-SVM. It conferred that proposed model achieves better accuracy except Yilmaz et al. study []. However, the proposed model and Yilmaz et al. study having almost similar results. Hence, it concluded that proposed model is an effective diagnostic model for predicting the diabetes.

Table 3.2: Comparative analysis of KHM-AW-SVM and thirty-one existing studies on diabetes using accuracy measure.

Sr. No.	Study	Algorithm	Accuracy (%)
1	Deng and Kasabov [106]	ESOM	78.4
2	Polat et al. [107]	LS-SVM,	78.21
3	Temurtas et al. [108]	MLNN with LM	79.62
4	Kayaer and Yıldırım [109]	GRNN	80.21
5	Carpenter and Markuzon [110]	ARTMAP-IC	81
6	Temurtas et al. [108]	MLNN with LM	82.37
7	Dogantekin et al. [111]	LDA-ANFIS	84.61
8	Bozkurt et al. [112]	DTDN	76
9	Yilmaz et al. [113]	Modified K-Means Clustering	93.71
11	Dogantekin et al. [111]	Various methods	59.5 and 77.7
12	Ramezani et al. [114]	LANFIS	88.05
13	Orkcu and Bal [115]	Real-coded Genetic Algorithm	77.6
14	Luukka [116]	Similarity Classifier + Feature Extraction	75.97
15	Isa et al. [117]	Clustered-Hybrid MLP	80.59
16	Ozcift and Gulten [118]	Rotation Forest Ensemble Classifier	74.47
17	Aslam et al. [119]	Genetic Programming+K-Nearest Neighbour	80.5
18	Seera and Lim [120]	Fuzzy Max-Min NN-CART Random Forest	78.39
19	Belle et al. [121]	Radial Basis Function Classifier	76.7
20	Wang et al. [122]	Improved Electro magnetism like Mechanism	77.21
21	Seera et al. [133]	Hybrid Fuzzy ARTMAP-CART model	87.64
22	Zhu et al. [85]	Multiple Factors Weighted Combination	93
23	Ding et al. [124]	Extreme Learning Machine	77.63
24	Mohapatra et al. [125]	Improved Cuckoo Search based ELM	78.5
25	Feng et al. [126]	Variable Coded Hierarchical Fuzzy Classification	79.17
26	Luukka [127]	Similarity Classifier using PCA and Entropy	75.82
27	Polat and Gunes [128]	Fuzzy-Artificial immune recognition system	84.42
28	Polat and Gunes [129]	PCA + ANFIS	89.47
29	Ghazavi and Liao [130]	Fuzzy Modelling with Selected Features	77.65
30	Polat et al. [107]	Generalized discriminant analysis Least square SVM	82.05
32	**Our Study**	**KhmAW-SVM (50-50 training and test set)**	**88.76**
33	**Our Study**	**KhmAW-SVM (10 cross fold)**	**92.38**

3.4 Summary

This chapter considers the accuracy issue of SVM classifier in context of diagnostic system. This issue of SVM is addressed through selection of optimal attributes for diagnosis process. So, in this work, KHM-AW method is developed for computing the weight of attributes and further, to determine the non-relevant attributes. The KHM method is used for measuring the label of data object. Further, a weight function is designed for computing the weight of attributes through weight coefficients and cluster centers. This method is integrated into SVM classifier for accurate prediction of diabetes diseases. The results are evaluated using 10-fold and 50-50% training-testing methods. Both all attributes and optimum attributes of diabetes are considered for evaluating the performance of proposed KHM-AW-SVM model. Results confirmed that attribute selection method provides advantage over SVM classifier. Furthermore, it is conferred that 10-fold method obtains better results than 50-50% training-testing method using all attributes and optimum attributes.

CHAPTER 4

ABC-DNN BASED DIAGNOSTIC MODEL FOR DIABETES

Feature selection (FS) is considered as an important activity for building an effective and promising diagnostic model. It extracts the more promising and effective features from the given medical dataset. Several researchers have developed the FS algorithm for selecting the optimum features for prediction task [8-11]. These techniques also reduce the computational cost and improve the accuracy rate. Further, these techniques integrate with ML techniques either in supervised or unsupervised manner and compute the weight of each feature. The weight of feature can decide the significance of feature in the prediction process also limit the number of input features for producing the good results through a diagnostic model [9-10]. This chapter considers the feature selection issue of medical dataset. This issue is handled through ABC based FS algorithm and aim of this algorithm is to compute optimal feature set for diabetes disease. Further, the selected features are fed to DNN technique for prediction task. Finally, a diagnostic model is developed using ABC based feature selection method and DNN technique, called ABC-DNN diagnostic model.

4.1 Chapter Contribution

The contributions of this chapter are highlighted as

1. Design an effective diagnostic model to improve the diagnosis accuracy of diabetes disease, called ABC-DNN.
2. Develop an ABC based FS algorithm to determine optimal feature for diagnosis of diabetes.
3. Applied DNN technique for accurate prediction of diabetes with reduced feature set.

4.2 Proposed ABC-DNN based Diagnostic Model

This section presents the proposed ABC-DNN based diagnostic model. The proposed model is the combination of ABC-FS algorithm and technique. The ABC-FS algorithm is discussed in the subsection 4.2.1. The aim of algorithm is to compute the optimal feature set for prediction task. While, subsection 4.2.2 illustrates the mechanism of describes DNN technique and this technique can be adopted as classifier. The schematic diagram of ABC-DNN is reported in Figure 4.1.

4.2.1 ABC Based Feature Selection (FS) Algorithm

The aim of ABC-FS algorithm is to choose the more optimum features of diabetes disease. The feature selection algorithm can be implemented either supervised or unsupervised paradigms. In this work, unsupervised paradigm can be considered for evaluating the optimal features. Hence, ABC-FS algorithm is implemented using unsupervised paradigms such as this algorithm initially computes optimal centroid for diabetes dataset and arranges the data into respective clusters using minimum distance criterion. Further, a weight function is adopted for computing the weight of each feature by considering the association of data with respective cluster centroids. On the other side, ABC is population based meta-heuristic algorithm derived through the behaviour of honey bees [131] and number of optimization problems are being solved by ABC algorithm till date [132-135]. The working of ABC is characterized through three types of bees-employed, onlooker and scout. Every bee is responsible to perform the specific task. The total number of bees can be defined in terms of employed bee, onlooker bee and scout bee. The ABC algorithm contains only one scout bee, while employed bee and onlooker bee are same. The global search of algorithm is described in terms of employed and onlooker bees, whereas, scout bee responsible for local search. The optimal solution is defined in terms of food source. The detailed steps of ABC-FS algorithm are given in Algorithm 4.1.

4.2.2 Deep Neural Network (DNN)

DNN can be characterized as a neural network with finite level of complexity. It consists of complicate mathematical model for processing of the information and contains multiple layer [6, 31]. DNN integrates both of feature selection and classification task itself to build good

Algorithm 4.1: ABC based Feature Selection Algorithm

Step 1: Initialize user defined parameters of the proposed algorithm such FS, limiting condition, clusters (K), maximum iterations, UB, LB

Determine the initial food source randomly within UB and LB constraints.

Evaluate the fitness of food source through employed bees.

Step 2: While (maximum iteration is not reached)

Step 3: For every employed bee, do following

Discover the position of new food source in the neighbouring of old food source using equation 4.1.

$$V_{i,j} = X_{i,j} + \emptyset_{i,j}\left(X_{i,j} - X_{k,j}\right) \tag{4.1}$$

Compute the fitness of the new generated food source using equation 4.2.

$$fit_i = \frac{1}{1 + f_i} \tag{4.2}$$

A greedy selection is made between the old and new food source positions and keep the best one

Step 4: Calculate the probability values for each food source using equation 4.3.

$$P_i = \frac{fit_i}{\sum_{i=1}^{SN} fit_i} \tag{4.3}$$

Step 5: For each onlooker bee, do following

If (rand () < P_i) /* P_i denotes probability of i^{th} food source */

Send the onlooker bee to exploit new food source in the neighbouring of old position using equation 4.4.

$$V_{i,j} = X_{i,j} + \emptyset_{i,j}\left(X_{i,j} - X_{k,j}\right) + \varphi_{i,j}\left(Y_j - X_{k,j}\right) \tag{4.4}$$

Compute the fitness of the new food source using equation 4.3.

A greedy selection is made between the old and new food source and keep the best one.

Else, i=i+1

Step 6: If (the food source is not upgraded further using limiting condition)

Send a scout bee to generate the new position of food source using equation 4.5.

$$X_{new} = X_{best} + rand[0,1](X_{best} - X_{curr}) \tag{4.5}$$

Step 7: Store the best solution and check the termination condition. If met, obtain the final optimized cluster centres. Otherwise Iteration = iteration + 1

Step 8: Compute the weight of each feature using equation 4.6.

$$f_i = \left(\sum_{i=1}^{d} \sum_{j=1}^{h} \sum_{k=1}^{k} \frac{X_{ih}}{C_{ik}}\right) \times \frac{1}{d} \tag{4.6}$$

Step 9: Select the features with maximum weight.

---

decision making. DNN obtains good attention from the research community in recent time. DNN can be described as an input layer (I) for input, hidden layer (L) for processing and an output layer (O) for final outcome. DNN is developed in python programming language using tf.contrib.learn.DNNClassifier library from Google. DNN having advantage over conventional neural network because it is designed using the wide set of trials and every trial modifies number of hidden layers, activation function, learning steps and number of neurons of DNN configuration. The accuracy of DNN is validated using the test set in every manual configuration. In this work, rectified linear unit activation function is used to generate all neuron layers for DNN classifier. The DNN classifiers is described through equation 4.7 and further, this equation is also defined the activation function.

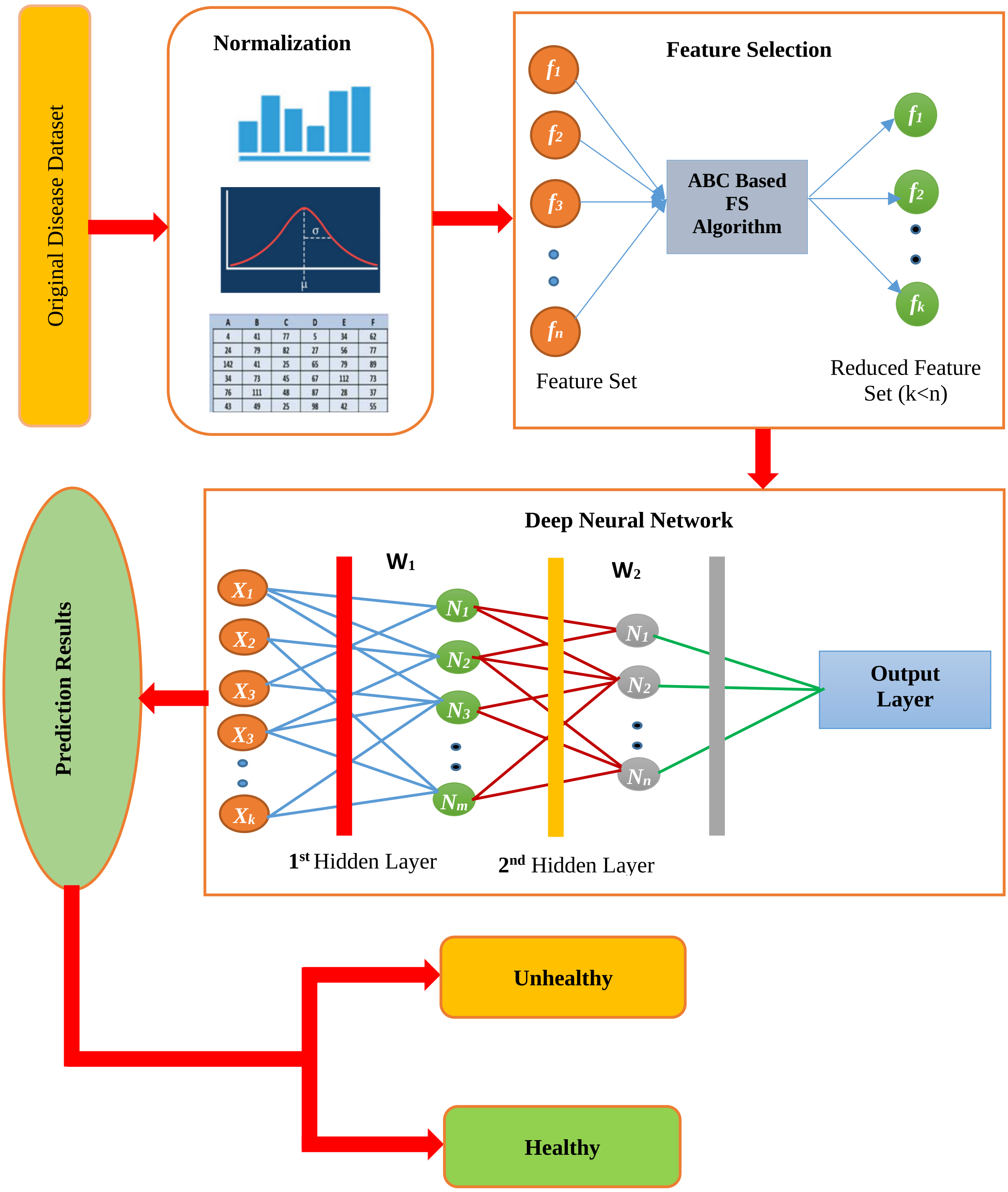


Figure 4.1: Block Diagram of the Proposed ABC-DNN based Diagnostic Model

$$F(x) = x' = \max(0, x) \tag{4.7}$$

In abovementioned equation, x denotes an input to a neuron, also called ramp function. Moreover, the activation function is associated with smooth approximation which is given in equation 4.8.

$$F(x) = \ln[1 + \exp(x)] \tag{4.8}$$

In prediction task, the new representation of input layer is illustrated through hidden layers and can be defined using equation 4.9.

$$x_{t+1} = m[w_t x_t + b_t] \tag{4.9}$$

x_{t+1} denotes the (t+1)th hidden layer, w_t denotes the weight of ith hidden layer, b_t denotes the bias of ith hidden layer and m denotes activation function .

4.3 Experimental Results and Discussion

This section demonstrates the simulation results of proposed ABC-DNN based diagnostic model. The performance of proposed model is evaluated using Pima indian diabetes dataset. Furthermore, 10 cross fold and 50-50 percent training and testing method is used to validate the performance of diagnostic model. The simulation results of proposed system are described in terms of average of the ten-independent run. The parameters tuning of the DNN method are taken same as reported in [31]. The prediction accuracy of the proposed model is described through accuracy, sensitivity, specificity, kappa, AUC and f-measure parameters.

Table 4.1 illustrates the simulation results of proposed ABC-DNN diagnostic model and DNN technique. The simulation results are evaluated using 10-cross fold and 50-50% training and testing techniques. It is observed that proposed diagnostic model obtains 94.74% accuracy rate using 10-fold method. Where, the accuracy rate using all features i.e. DNN technique is 82.67%. Hence, it is stated that proposed ABC based feature selection method improves the accuracy rate of DNN technique. On the analysis of 50-50% training-testing technique, it is also seen that proposed ABC-DNN diagnostic model achieves higher accuracy rate as compared to DNN technique. It is observed that ABC based feature selection method improves the accuracy rate of DNN upto twelve percent. It is also stated that proposed diagnostic model achieves higher sensitivity, specificity, f-measure, kappa and AUC rate as compared to DNN technique. Furthermore, it is noted that kappa statistics is significantly improved using ABC-

DNN technique. It is also revealed that there is significant difference between the performances of 10-fold and 50-50% techniques in terms of ABC-DNN model.

Table 4.1: Simulation results of DNN and ABC-DNN on diabetes dataset.

Features	Parameters	10 cross fold	50-50% training and test
Using All Features (DNN)	ACC	82.67 ± 4.53	81.06 ± 4.87
	Sensitivity	88.93± 5.14	87.41 ± 4.74
	Specificity	78.58 ± 6.13	74.59 ± 5.81
	f-measure	78.12	75.67
	Kappa	71.23	68.43
	AUC	81.42	79.08
Using Attribute Weighting (ABC-DNN)	ACC	94.74 ± 2.56	91.61 ± 3.46
	Sensitivity	95.52 ± 4.18	93.16 ± 3.94
	Specificity	92.06 ± 2.35	90.04 ± 4.23
	f-measure	89.31	85.94
	Kappa	91.28	87.56
	AUC	92.36	88.27

The experimental results of proposed ABC-DNN based diagnostic model is also compared with existing diabetes prediction model and several machine learning approaches. Table 4.2 presents the experimental results of proposed diagnostic model and other existing diabetes models using accuracy parameter. It is revealed that ABC-DNN model obtains higher accuracy rate as compared to rest of existing studies. Hence, it is stated that proposed ABC-DNN based diagnostic model is one of the efficient and effective method for computer-aided diagnostic system. This model also improves the accuracy rate of diabetes disease. Further, it is noticed that proposed diagnostic model obtains higher accuracy rate using 10-fold technique rather than 50-50% training-testing technique.

Table 4.2: Performance comparison of proposed ABC-DNN method and previous studies reported in literature

Sr. No.	Study	Algorithm	Accuracy (%)
1	Deng and Kasabov [106]	ESOM	78.4
2	Polat et al. [107]	LS-SVM,	78.21
3	Temurtas et al. [108]	MLNN with LM	79.62
4	Kayaer and Yıldırım [109]	GRNN	80.21
5	Carpenter and Markuzon [110]	ARTMAP-IC	81
6	Temurtas et al. [108]	MLNN with LM	82.37
7	Dogantekin et al. [111]	LDA-ANFIS	84.61
8	Bozkurt et al. [112]	DTDN	76
9	Yilmaz et al. [113]	Modified K-Means Clustering	93.71
11	Dogantekin et al. [111]	Various methods	59.5 and 77.7
12	Ramezani et al. [114]	LANFIS	88.05
13	Orkcu and Bal [115]	Real-coded Genetic Algorithm	77.6
14	Luukka [116]	Similarity Classifier + Feature Extraction	75.97
15	Isa et al. [117]	Clustered-Hybrid MLP	80.59
16	Ozcift and Gulten [118]	Rotation Forest Ensemble Classifier	74.47
17	Aslam et al. [119]	Genetic Programming+K-Nearest Neighbour	80.5
18	Seera and Lim [120]	Fuzzy Max-Min NN-CART Random Forest	78.39
19	Belle et al. [121]	Radial Basis Function Classifier	76.7
20	Wang et al. [122]	Improved Electro magnetism like Mechanism	77.21
21	Seera et al. [133]	Hybrid Fuzzy ARTMAP-CART model	87.64
22	Zhu et al. [85]	Multiple Factors Weighted Combination	93
23	Ding et al. [124]	Extreme Learning Machine	77.63
24	Mohapatra et al. [125]	Improved Cuckoo Search based ELM	78.5
25	Feng et al. [126]	Variable Coded Hierarchical Fuzzy Classification	79.17
26	Luukka [127]	Similarity Classifier using PCA and Entropy	75.82
27	Polat and Gunes [128]	Fuzzy-Artificial immune recognition system	84.42
28	Polat and Gunes [129]	PCA + ANFIS	89.47
29	Ghazavi and Liao [130]	Fuzzy Modelling with Selected Features	77.65
30	Polat et al. [107]	Generalized discriminant analysis Least square SVM	82.05
31	**Our Study**	**ABC-DNN (50-50 training and test set)**	**91.61**
32	**Our Study**	**ABC-DNN (10 cross fold)**	**94.74**

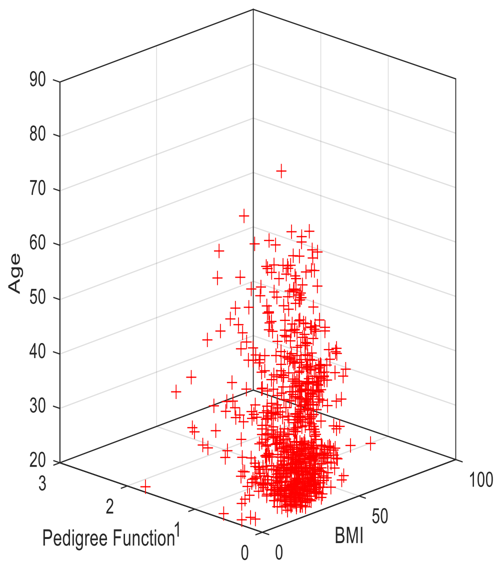


Fig. 4.2(a)

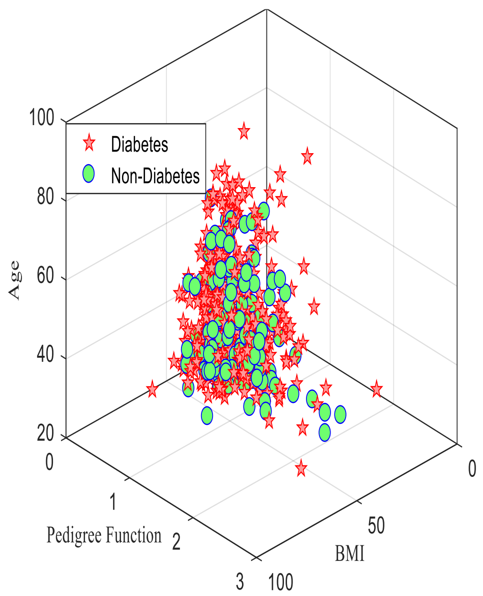


Fig. 4.2(b)

Figure 4.2(a and b): (a) Illustrate the distribution of data and (b) shows diabetes prediction using ABC-DNN method.

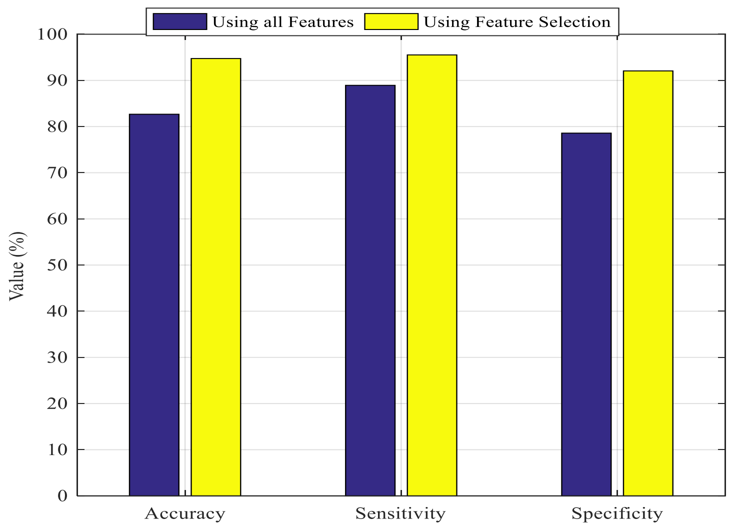
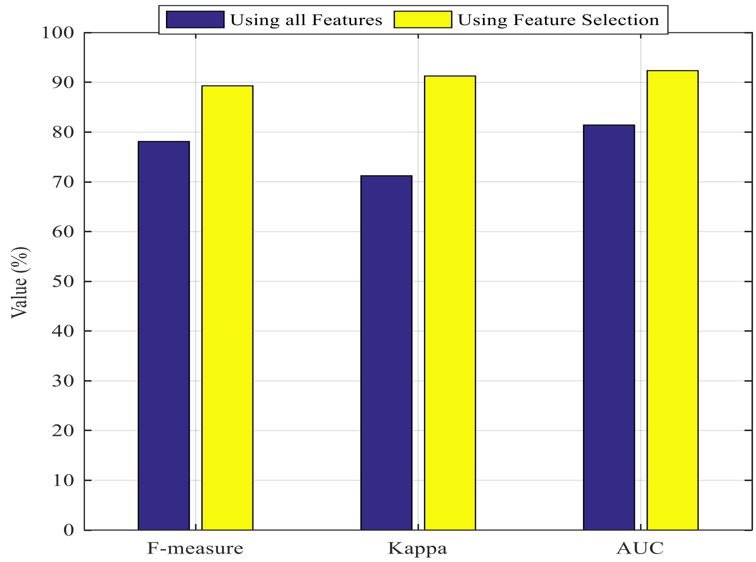


Figure 4.3: Simulation results of DNN and proposed ABC-DNN based diagnostic model using accuracy, sensitivity and specificity parameters (10-fold method)



Figures 4.4: Simulation results of DNN and proposed ABC-DNN based diagnostic model using F-measure, Kappa and AUC (10-fold technique)

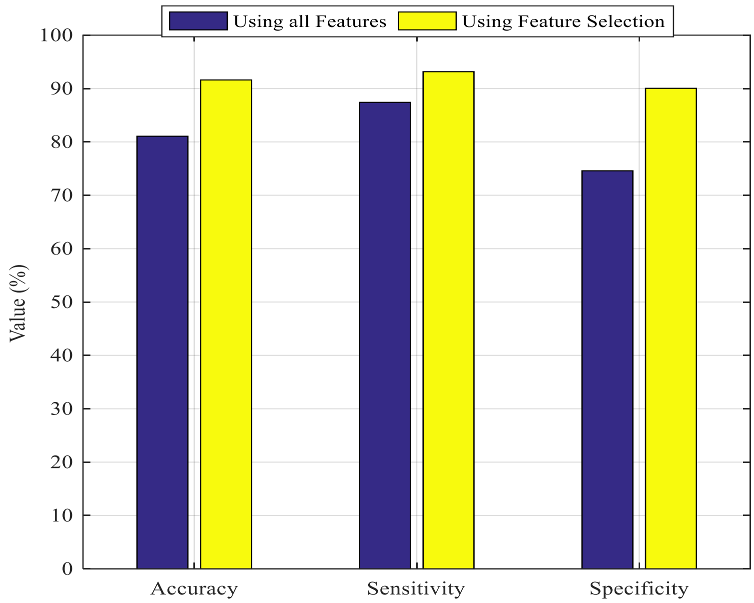
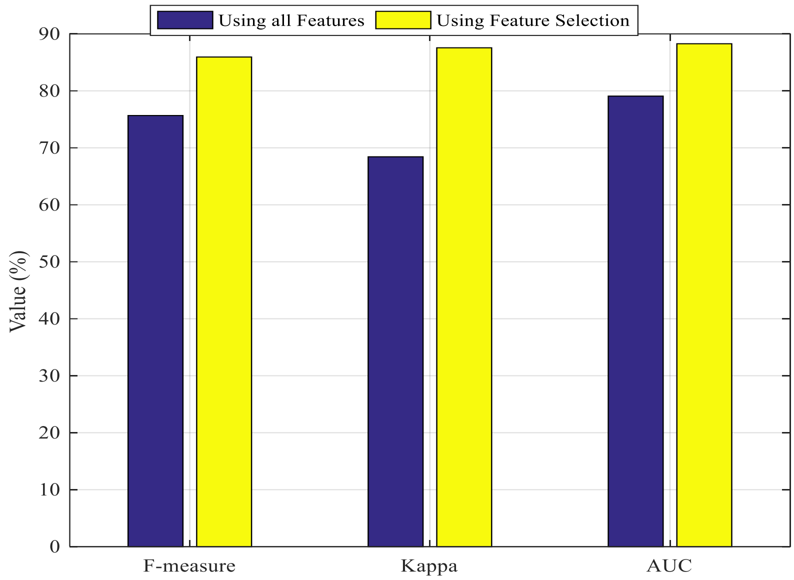


Figure 4.5: Simulation results of DNN and proposed ABC-DNN based diagnostic model using accuracy, sensitivity and specificity parameters (50-50% training-testing method)



Figures 4.6: Simulation results of DNN and proposed ABC-DNN based diagnostic model using F-measure, Kappa and AUC (50-50% training-testing method)

Figure 4.2(a) illustrates the data instances of diabetes dataset using age, pedigree function and BMI features. While, Figure 4.2(b) denotes the prediction of diabetes dataset in two classes i.e. Diabetes and Non-Diabetes. The prediction of diabetes and non-diabetes patients is done through the proposed ABC-DNN based diagnostic model. Figures 4.3-4.4 demonstrate the experimental results of the ABC-DNN based diagnostic model and DNN technique using 10-fold method. Figure 4.3 shows the experimental results of ABC-DNN model and DNN technique using accuracy, sensitivity and specificity parameters. Whereas, Figure 4.4 depicts the simulation results using kappa, AUC and F-measure parameters. It is seen that proposed ABC-DNN based model achieves better quality results than DNN technique. Figures 4.5-4.6 show the experimental result of ABC-DNN model and DNN technique using 50-50% training-testing method. It is noticed that integration of ABC based feature selection method improves the performance of DNN significantly. Hence, it is stated that ABC-DNN based diagnostic model outperforms than DNN technique.

4.4 Summary

This chapter presents an ABC-DNN based diagnostic model for the diagnosis and prediction of diabetes disease. The proposed diagnostic model is designed using ABC based feature selection method and deep neural network technique. In proposed model, ABC based feature selection method is used to determine the relevant features of diabetes disease. Further, the DNN technique is adopted for diagnosis and prediction of diabetes disease. The performance of proposed diagnostic model is evaluated using Pima Indian Diabetes dataset. The different performance measures like accuracy, sensitivity, specificity, AUC, F-measure and Kappa are considered to assess the performance of ABC-DNN based diagnostic model. Furthermore, experimental results are evaluated using 10-fold and 50-50% training-testing techniques. It is observed that ABC-DNN based diagnostic model provides better results than DNN method. The experimental results of ABC-DNN model is also compared with thirty-one existing diabetes studies. It also revealed that proposed diagnostic model achieves higher accuracy rate than existing studies. Furthermore, it is noticed that 10-fold method is more suitable than 50-50% training-testing method.

CHAPTER 5

HYBRID DIABETES DISEASE PREDICTION FRAMEWORK

The earlier detection and management of diabetes is an important activity for reducing the number of deaths due to diabetes. It is noticed that late diagnosis increases the number of deaths. Hence, the integration of data mining and information technology can be utilized as suitable cutting-edge technological solution for earlier detection and diagnosis of diabetes. Data mining is the branch of computer science which concerns with the exploration of hidden information from the large databases and this unexplored information can be used in medical diagnosis and decision-making. Large number of decision and diagnostic models are presented in literature for effective management and detection of diabetes. [7-11]. But, analysis of diabetes data is challenging task due to nonlinearity, non-normal, structured correlation, and complex of nature medical data [12]. It is observed that ML-based systems dominating in the medical healthcare field [13–16] and widely for medical imaging such as stroke, coronary artery disease, and cancer [17-19]. Moreover, machine learning techniques can be used for feature selection as well as classifiers. These techniques also help to medical practitioner for accurate diagnosis and risk stratification of diabetes. Hence, this chapter addresses the missing value imputation and attribute selection issues related to diabetes dataset.

5.1 Chapter Contribution

This chapter aims to develop a hybrid diabetes disease framework for accurate prediction of diabetes disease. The proposed framework handles the two well-known issues of medical data such as missing value imputation and presence of outlier.

1. Develop a K-Mean++ based data imputation technique for imputing missing value.
2. Design an ABC based outlier detection technique for addressing outliers in data.

3. Adopt Least Square-Support Vector Machine for diagnosis of diabetes.

5.2 Proposed Hybrid Diabetes Prediction Framework

This section describes the working of the proposed diabetes prediction framework. The proposed framework contains K-Mean^{++} based missing value imputation method, ABC based outlier detection method, and least square- support vector machine (LS-SVM) as classification technique. Figure 5.1 illustrates the working of proposed diabetes prediction framework.

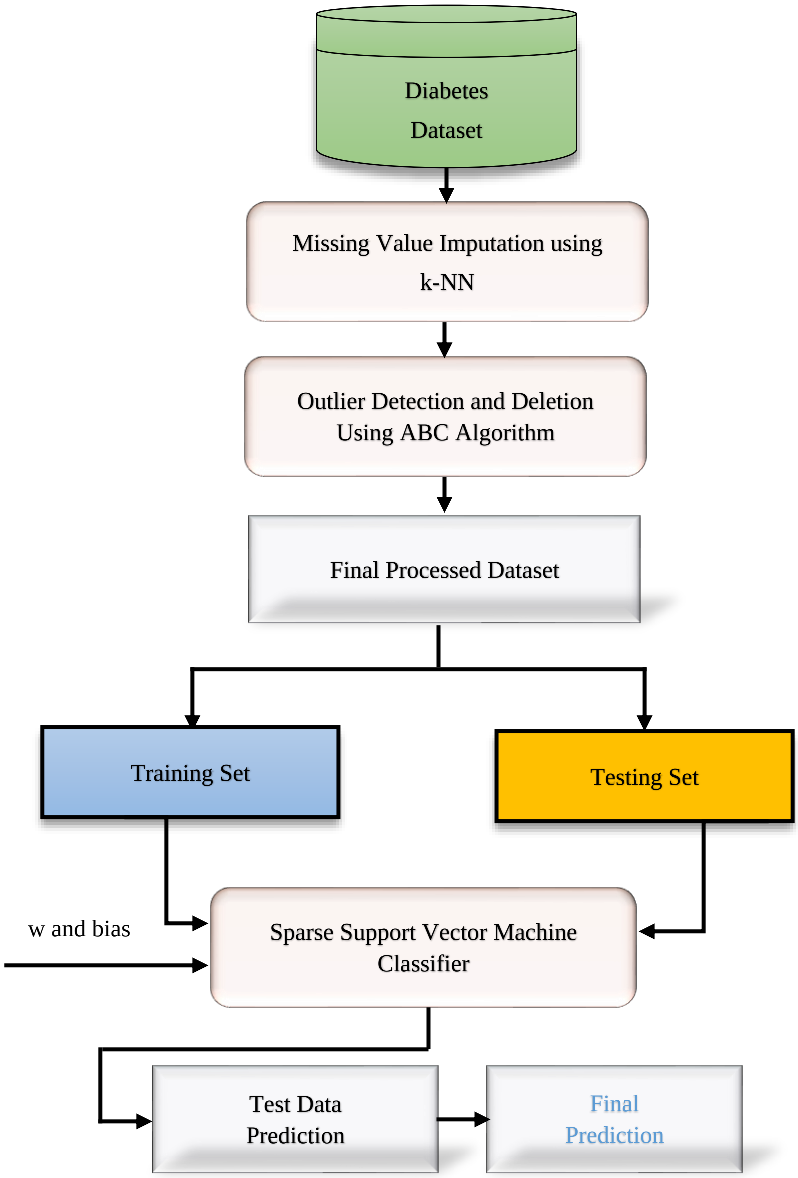


Figure 5.1: Process of proposed diabetes prediction framework

5.3 Pre-processing of Data

The pre-processing of data is an important phase of every prediction system. Similarly, the proposed diabetes framework also contains the data pre-processing phase. The aim of this phase is to convert the raw data into pre-processed data. This phase includes data imputation and outlier detection techniques. The description of these techniques is discussed into subsections 5.3.1.1 and 5.3.1.2

5.3.1 K-Mean^{++} Data Imputation Technique

The real-world databases contain missing information and incomplete data sometimes. Several reasons are responsible for missing information and incomplete data. These reasons can be summarized as errors in data collection, data entry methods, improper measurements, equipment malfunctioning etc. The presence of missing values in databases can lead to occurrence of many problems during knowledge discovery process. These problems are summarized as lack of efficacy, difficulty for managing the data and data analysis. These missing values can also lead to bias decision as data is incomplete. Hence, exact prediction is not possible from such data. In literature, missing values are handled through either data tolerance techniques or data imputation techniques [136-138]. The data tolerance technique can be described through data mining algorithms such as clustering, classification and feature selection, but these techniques are not computing the missing values. While, data imputation techniques compute the missing values and fill these values in dataset to make it complete. Table 5.1 illustrates a snapshot of Pima Indian Diabetes Dataset with missing values. Table 5.1 describes the snapshot of pima Indian diabetes dataset in which missing value is denoted through “*” symbol.

Table 5.1: Snapshot of missing value in Pima Indian Diabetes Dataset.

Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree	Age
3	116	74	15	105	26.3	0.107	24
*	117	66	31	188	30.8	0.493	22
0	*	65	*	*	24.6	0.66	31
2	122	60	18	106	*	0.717	22
*	107	76	*	*	45.3	0.686	24
1	86	66	52	65	*	0.917	29
6	91	*	*	*	29.8	0.501	31
1	*	56	30	56	33.3	1.251	24
4	132	*	*	*	32.9	0.302	23
*	105	90	*	*	*	0.197	46
0	57	60	*	*	21.7	0.735	67

The analysis of diabetes dataset confirmed that 432 data instance consists of missing values. A total 763 missing values are presented into diabetes dataset. Out of 8 features, 6 features contain the missing values. These features are pregnant, plasma glucose, diastolic BP, SFT, serum insulin and BMI and quantity of missing values are 111, 5, 35, 227, 374 and 11 respectively. Its stated that serum insulin attribute having 48.6% missing value, while, SFT attribute contains 29.5% missing values. Due to such extent of missing values, it is not possible that unbiased can be taken. Hence, to compute the missing value for diabetes dataset, K-Mean^{++} data imputation method is proposed and aim of this method is to compute the missing value in effective manner. The steps of K-Mean^{++} data imputation method is listed in Algorithm 5.1.

5.3.2 ABC Based Outlier Detection Technique

In data analysis process, outlier detection is an important activity and accuracy can be improved by removing the outlier from the dataset. Through literature, it is revealed that clustering algorithms are widely adopted for detecting and removing the outlier in datasets [139]. These methods determine the outliers using distance function. Furthermore, the general approach stated that an outlier is described through small and low-density clusters or not associated with any clusters. whereas, normal data is described through large and dense clusters. Several different methods are applied for detecting outliers in datasets. These methods are summarized as statistical methods [140], clustering based methods [141] and distance-based methods [142] etc.

Algorithm 5.1: Steps of K-Means^{++} data imputation method	
Input:	Diabetes dataset (D) and number of clusters (K)
Output:	Missing value computation and complete processed dataset
Step 1:	Load the dataset (D) and determine initial cluster center $(c_1 \in K)$ from dataset(D) in uniform order.
Step 2:	Compute the next cluster center (c_i) such that $(c_i) = x' \in D$ using probability function mentioned in equation 5.1. $$c_i = \frac{dist(x')^2}{\sum_{x \in D} dist(x)^2} \quad (5.1)$$ dist(x) denotes the shortest distance between data (x) to nearest center that randomly chosen.
Step 3:	Repeat the step 2, until all cluster centers are not determined.
Step 4:	Compute the Euclidean distance between cluster centers $(c_i \in K)$ and each data (x) presented in dataset (D).
Step 5:	Allocate the data(x) to clusters $(c_i \in K)$ with minimum Euclidean distance.
Step 6:	Recompute the new clusters using equation 5.2. $$c_{i,new} = (1/c_i) \sum_{k=1}^{c_i} x_i \quad (5.2)$$

```
Step 7:   Repeat the steps 4-6, until, there is no change in
          data allocation between clusters. Otherwise obtain
          final centroid.
Step 8:   Compute the arrangement of data instances as per
          final centroid.
Step 9:   Data instances within cluster acted as nearest
          neighbor for each other.
Step 10:  Compute the mean value of each nearest neighbor
          (mean value of each clusters) and replaced the
          missing value through mean value (in case of
          numeric attributes)
Step 11:  Obtain the complete processed dataset.
```

The statistical methods consider the distribution of data for detecting the outliers. The data that are deviated from given standard distribution can be interpreted as outliers. Such techniques can be worked efficiently, if distribution of data is known in advance, but it is not possible in case of high dimensional data and large datasets. In clustering-based method, behaviour of data is characterized by constructing a clustering model. Further, the distance between data and respective cluster centroid can be used to determine the outliers. It is assumed that if a data is far away from respective cluster centroid using a significant distance, it is outlier. It is also reported that clustering-based methods can determine outlier in effective manner. In distance-based methods, the distance between data and other data within dataset can be used for detecting the outlier and also known as local outlier factor. These methods compute the distance through density estimation and k-nearest neighbour. The above discussion conclude that clustering-based algorithm is more suitable for outlier detection. Hence, this work also explores the efficacy of clustering-based algorithm for detecting the outliers. To achieve the same, an artificial bee colony (ABC) based clustering algorithm [135] is adopted to determine the outliers in diabetes dataset. ABC is popular technique that is inspired through bee behaviour and obtained state of art results for many optimization problems [132-133].

Algorithm 5.2: Steps of ABC based clustering algorithm for outlier detection	
Input:	Diabetes dataset (D), Significant distance (SD) and number of clusters (K)
Output:	K-clusters and Outlier set(O)
Step 1:	Load the diabetes dataset (D) and user defined parameter like food sources (lusters), limit, maximum iteration, SD, colony size, lower limit and upper limit.
Step 2:	Determine initial food source (cluster centroids) in the range of lower limit and upper limit such as $(F_k \in K) \in D$ in random nature.
Step 3:	Compute the objective function (Euclidean distance) using initial food source positions (clusters centroid) and data presented in dataset(D).
Step 4:	Allocate the Data to food sources (K) according the minimum objective function value.
Step 5:	Allocate the data(x) to clusters $(c_i \in K)$ with minimum Euclidean distance.
Step 6:	Move employed bees for searching the foods in nearby area.
Step 7:	While (iteration $\leq$ maximum iteraion)
Step 8:	Execute EB Phase For i: 1 to K • Explore new food sources near to old food sources using equation 5.3.

$$V_{i,j} = X_{i,j} + \emptyset_{i,j}\left(X_{i,j} - X_{k,j}\right) \tag{5.3}$$

- Determine fitness of new food sources using equations 5.4-5.5.

$$\mathrm{fit}_i = \frac{1}{1 + f_i} \tag{5.4}$$

$$f_i = \frac{1}{n_j} \sum_{i=1}^{n_j} d\left(X_j, C_i\right) \tag{5.5}$$

- Compute best one through greedy selection process and store the best in memory.

End for

Step 9: Probability of food sources is computed through equation 5.6.

$$P_i = \frac{\mathrm{fit}_i}{\sum_{i=1}^{SN} \mathrm{fit}_i} \tag{5.6}$$

Step 10: Execute OB Phase

For i = 1: K

If (rand () < Pi)

- Explore new food sources near to old food sources using equation 5.7.

$$V_{i,j} = X_{i,j} + \emptyset_{i,j}\left(X_{i,j} - X_{k,j}\right) + \varphi_{i,j}\left(Y_j - X_{k,j}\right) \tag{5.7}$$

- Determine fitness of new food sources using equations 4-5.
- Compute best one through greedy selection process and store the best in memory

End For

Step 11: Execute SB Phase

If (food source is not improved wrt to limit)

- Generate the new food source using equation 5.8.

$$X_{new} = X_{best} + \mathrm{rand}[0,1](X_{best} - X_{curr}) \tag{5.8}$$

End If

Iteration=iteration+1.

```
Step 12:  Memorize the best solution and Iteration = iteration
               + 1
          End While
Step 13:  Obtain the optimal solution and arranged the data
          into clusters.
Step 14:  For i= 1 to K
            • Compute the distance between data and respective
              cluster centroids.
            If (Computed Distance>SD)
            • Marked Data as outlier and store into outlier
              set (O)
            Else
            • Marked data as normal data and put in data
              matrix of clusters (K)
          End for
Step 15:  Return K-clusters and Outlier set(O)
```

5.4 Processed Data

As the results of K-Mean^{++} based missing data imputation and ABC based outlier detection techniques, the pre-processed data is converted into processed data. The K-Mean^{++} technique effectively handles the missing data problem of diabetes dataset and computes the optimal values for missing data. Further, outliers presented in dataset are detected through ABC based outlier detection technique and these outliers are removed from the dataset. Hence, these techniques efficiently convert the pre-processed data into processed data and this data can fed to classifier for prediction task.

5.5 Prediction using LS-Support Vector Machine (LS-SVM)

This section describes the predictive task using support vector machine (SVM) on diabetes dataset. SVM is robust classifier and widely adopted for predictive tasks [29,33]. However, overfitting problem is associated with SVM classifier and this issue is ridden off through minimization of structural risk. The other variant of SVM variant, called LS-SVM is presented in [143]. The inequal constraints are translated into equal constraints through minimized the squared error instead of negative error and LS-SVM describes through linear equations instead of quadratic programming. Thus, margin errors and least squares errors can be minimized at same time, in turn prediction accuracy can be improved.

LS-SVM variant can be formulated for binary calcification as. For the given input or training sample (TS) can be described as $\mathrm{TS} : (\mathrm{X_i}, \mathrm{T_j})$ where $\mathrm{i} = 1, 2, 3, \dots \dots, \mathrm{N}$ and $\mathrm{T_j} \in \{-1. +1\}$ and decision function of LS-SVM computed using given equation 5.9.

$$\mathrm{F(X)} = w^T \times \xi(\mathrm{X}) + \mathrm{c} \tag{5.9}$$

In equation 9, w denotes the weight and c denotes the bias of LS-SVM. $\xi(.)$ can be interpreted as non-linear function that can maps input into high dimensional space. The LS-SVM with objective function and equality constraints can be descried using equation 5.10.

$\mathrm{Min\, S\, (w, c, a_i)} =$

$$\frac{1}{2}\mathrm{w^T w} \times \frac{1}{\gamma}\sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{a_i^2} \ \mathrm{S.T.} \quad \mathrm{w^T} \times \xi(\mathrm{X}_i) + \mathrm{c} = \ \mathrm{b_i} - \ \mathrm{a_i} \ \text{ where } \mathrm{i} = 1{,}2, \dots, \mathrm{N} \tag{5.10}$$

In equation 5.10, $\mathrm{a_i}$ denotes a slack variable, X_i, and γ describes as penalty parameter of LS-SVM. The flowchart of the LS-SVM is given Figure 5.2. Furthermore, radial basis function

(RBF) can be used as kernel function for LS-SVM due to its simplicity and effectiveness [36]. It also improves the performance of LS-SVM in significant manner with reduced computational cost. Hence, this work also considers the RBF kernel function for LS-SVM and it is explained in equation 5.11.

$$\mathrm{k}(\mathrm{X_i} - \mathrm{X_j}) = \exp(-\delta\|\mathrm{X_i} - \mathrm{X_j}\|) \tag{5.11}$$

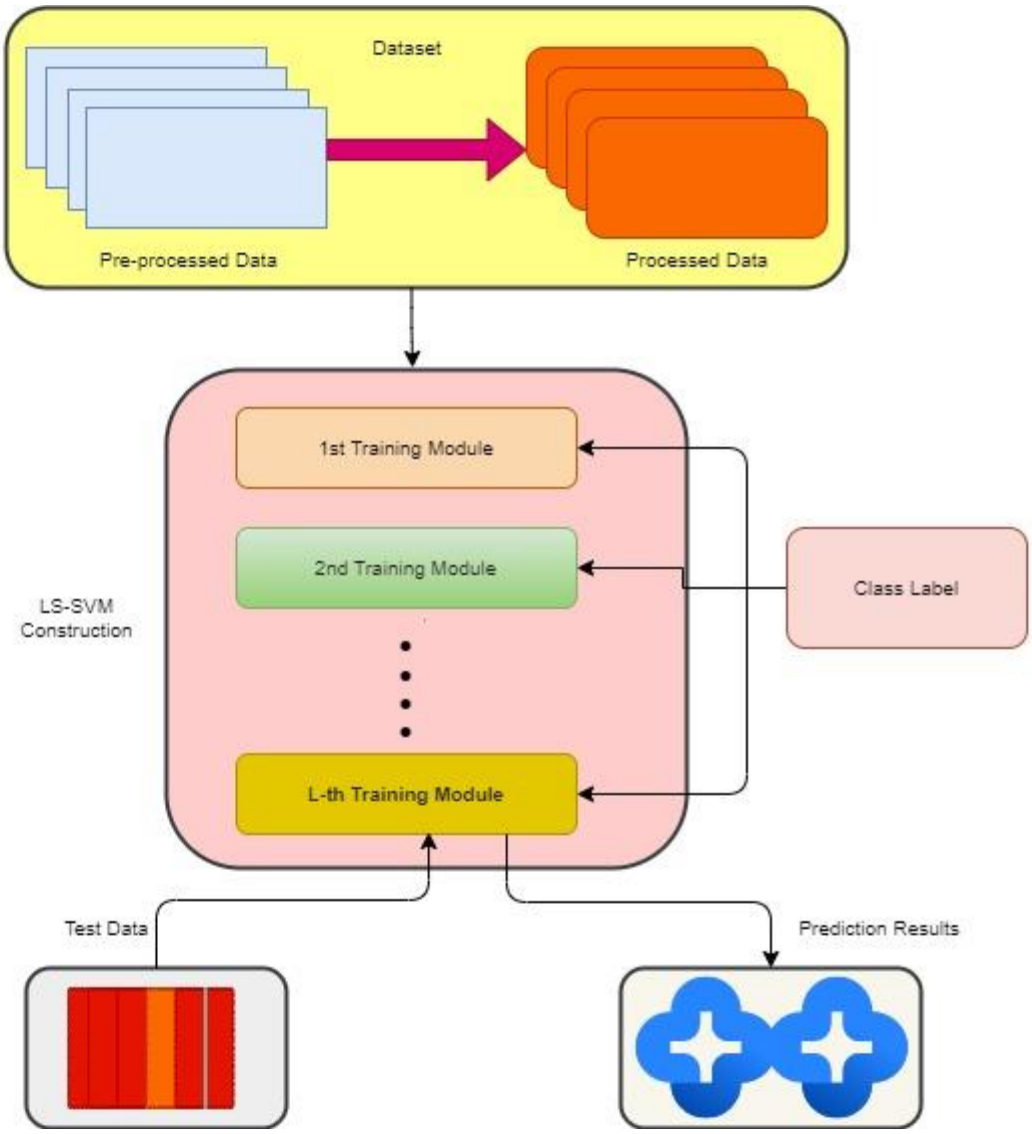


Figure 5.2: Flowchart of LS-SVM for diagnosis of diabetes disease

5.6 Experimental Results and Discussion

This section validates the performance of proposed hybrid diabetes frameworks using Pima Indian Diabetes dataset. The proposed framework consists of three different techniques for accurate diagnosis of diabetes. These methods are K-Mean++ based missing value imputation, ABC based outlier detection and LS-SVM techniques. The K-Mean++ technique is applied to overcome the missing values problem. A total 432 data instances of diabetes dataset contain missing values and these missing values are presented in six attributes out of total eight attributes of diabetes disease dataset. In general, total 763 missing values are presented in diabetes dataset and most of researchers replace these values by "0". In this work, the missing value problem of diabetes dataset is handled through K-Mean++ based missing value imputation technique and this technique compute the more optimum values for the same. The detailed description of this technique is given in subsection 5.3.1.1. The outlier is also affected the performance of predictive classifiers. Hence, this work also considers this problem and proposed an ABC based outlier detection technique. The aim of this technique is to improve the predictive accuracy of classifiers. Further, the patterns are predicted using LS-SVM classifiers. The performance of proposed hybrid diabetes framework is evaluated using accuracy, sensitivity, specificity, kappa and area under curve (AUC) parameters. The above discussed parameters are derived through confusion matrix.

Table 5.2 presents the clustering results of K-Mean^{++} data imputation technique. The diabetes data set contains 768 data instances and Cluster_1 contains 268 data instances, whereas Cluster_0 contains 500 data instances. The aim of imputation technique is to allocate the data instances to Cluster_1 and Cluster_0. Cluster_1 corresponds to diabetes positive and cluster_0 corresponds to diabetes negative. The above discussed imputation technique allocates 207 data instances correctly to Cluster_1 and 394 data instances correctly to Cluster_0. A total 167 data instances are incorrectly classified; 61 incorrectly data instances belong to Cluster_1 and 106 data instances belong to Cluster_0. In turn, these data instance remove form the diabetes dataset and the missing values associated with rest of data instances are computed using an average distance of data instance to respective cluster centroids. The diabetes data set contains total 601 data instances after K-Mean++ based data imputation technique.

Table 5.3 illustrates the simulation results of proposed K-Mean^{++} data imputation technique and other missing value data imputation techniques presented in literatures as percentage of incorrectly classified instances. These techniques are FKMI, KMI, KNNI, LSSI, MC, SVDI, SVMI and WKNN. Table 3 showed that proposed K-Mean++ data imputation technique achieves lower error rate i.e. 21.74 (in terms of incorrectly classified data instances) as compared to other techniques. SVDI technique obtains higher error rate i.e. 41.95 among all techniques. It concluded that K-Mean++ data imputation technique allocates data instances to respective clusters in more efficient way, while SVDI exhibits lower efficacy.

Table 5.2: Clustering results of K-Mean^{++} missing value data imputation technique.

Cluster Attribute	Total Instances	Correctly Classified	Incorrectly Classified
Cluster_1 (+)	268	207	61
Cluster_0 (-)	500	394	106

Table 5.3: Results of missing values imputation techniques on diabetes dataset (601 data instance after K-Mean^{++})

S. No.	Missing Value Imputation Technique	Error Rate (%)
1	K-Mean++	21.74
2	FKMI	30.36
3	KMI	27.68
4	KNNI	25.72
5	LSSI	31.09
6	MC	31.58
7	SVDI	41.95
8	SVMI	30.24
9	WKNN	26.83

This work also explores the efficacy of ABC technique for outlier detection. The results of ABC based outlier detection technique and other popular techniques are presented in Table 5.4. These techniques are K-NN, Isolation Forest, Histogram, LOCI, Feature Bagging, K-Mean, K-Mean++, Percentiles, and MCD. Simulation results showed that proposed ABC technique detects more outlier i.e. 7.82% in comparison to other technique. ABC technique determines 38 data instances as outlier and these outliers are removed from the diabetes dataset. It also found that Percentiles technique detects lower outlier i.e., 4.33 as compared to other techniques.

Table 5.4: Results of different outlier detection techniques on diabetes dataset

S. No.	Outlier Detection Technique	Outlier Detected (%)	Data Instances
1	K-NN	6.32	38
2	Isolation Forest	4.82	29
3	Histogram	5.32	32
4	LOCI	4.99	30
5	Feature Bagging	5.67	34
6	K-Mean	6.15	37
7	K-Mean++	6.65	40
8	Percentiles	4.33	26
9	MCD	5.83	35
10	ABC	7.82	47

After removing the outliers, a total 554 data instances are left in diabetes dataset, called processed dataset. Finally, the LS-SVM classifier is applied to predict the final outcomes either diabetes positive or negative patients. The outcome of LS-SVM is mentioned using confusion matrix. Table 5.5 shows the confusion matrix obtained through proposed hybrid diabetes framework. The proposed framework determines 535 data instances correctly and 19 data instances incorrectly. Furthermore, the performance of proposed hybrid diabetes framework is also evaluated using accuracy, sensitivity, specificity, kappa and AUC parameters. All these parameters are derived from confusion matrix.

Table 5.5: Confusion matrix (total instance 554) obtained through proposed hybrid diabetes framework.

(+)	(-)	Classified as
169 (TP)	12 (FN)	(+) → Diabetes Positive
7 (FP)	366 (TN)	(-) → Diabetes Negative

Table 5.6 presents the simulation results of proposed framework using accuracy, sensitivity, specificity, kappa and AUC parameter. The results are compared with LS-SVM, (K-Mean^{++}+LS-SVM), (ABC+SVM) techniques. Its revealed that proposed framework obtains higher accuracy i.e. 96.57 as compared to other compared techniques. The proposed framework

achieves higher sensitivity and specificity rates i.e. 93.97 and 98.97 than other techniques. The kappa and AUC parameters also prove the effectiveness of the proposed hybrid diabetes framework than other technique. Hence, it is said that proposed hybrid diabetes framework accurately predicts the diabetes affected patients.

Table 5.6: Results of LS-SVM, (K-Mean^{++}+LS-SVM), (ABC+SVM) and Proposed Framework (K-Mean^{++} +ABC+LS-SVM) techniques

Parameter	Technique			
	LS-SVM	(K-Mean^{++}+LS-SVM)	(ABC+SVM)	Proposed Framework
Accuracy	84.24	89.37	87.89	96.57
Sensitivity	81.56	88.04	86.46	93.37
Specificity	83.78	90.64	88.21	98.12
Kappa	86.52	89.83	89.34	98.17
Area Under Curve	82.28	87.91	85.43	95.43

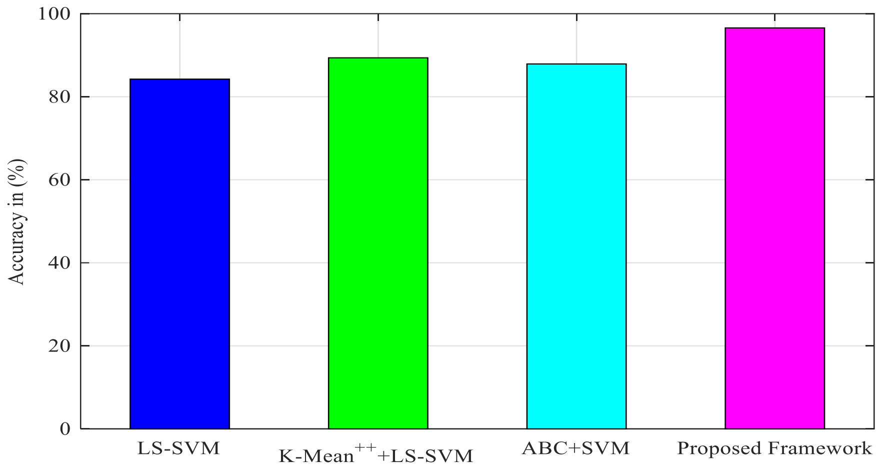


Figure 5.3: Comparison of accuracy rate of LS-SVM, (K-Mean^{++}+LS-SVM), (ABC+SVM) and proposed framework

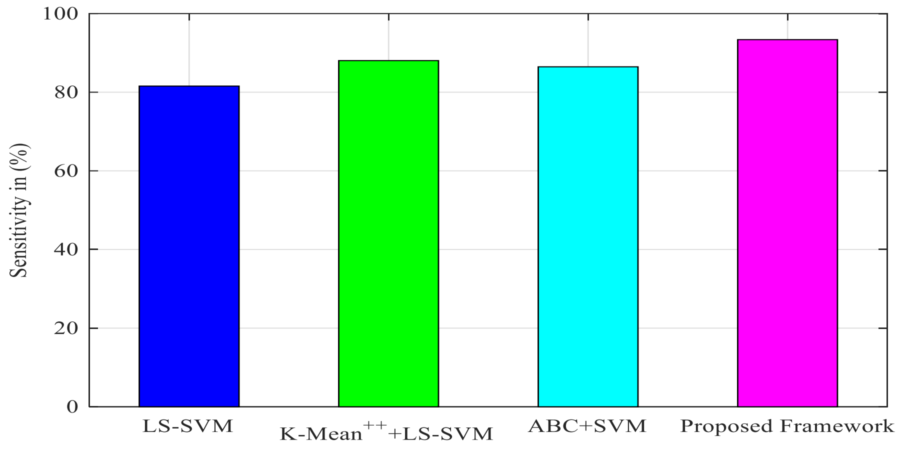


Figure 5.4: Comparison of sensitivity rate of LS-SVM, (K-Mean^{++}+LS-SVM), (ABC+SVM) and proposed framework

Figures 5.3-5.4 demonstrates the prediction results of LS-SVM, (K-Mean^{++}+LS-SVM), (ABC+SVM) and proposed diabetes framework in terms of accuracy and sensitivity rates. It is analysed that proposed framework obtains better results for both of parameters as compared to other techniques. It is also observed that LS-SVM exhibits lower performance for diabetes dataset in terms of accuracy and sensitivity rates. It is also stated that outlier detection and data imputation techniques improve the accuracy and sensitivity rates in significant manner.

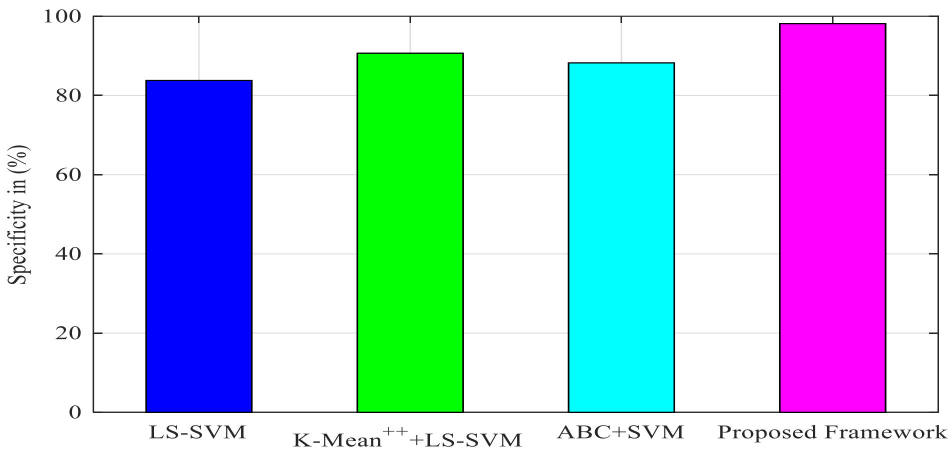


Figure 5.5: Comparison of specificity rate of LS-SVM, (K-Mean^{++}+LS-SVM), (ABC+SVM) and proposed framework

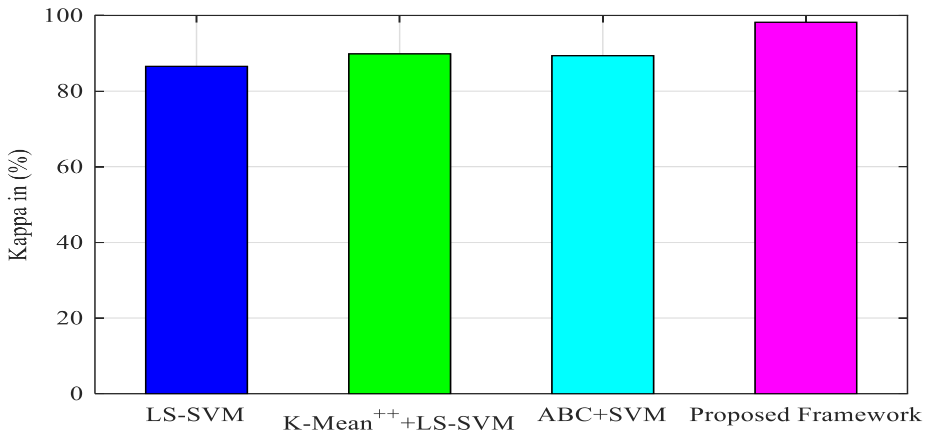


Figure 5.6: Comparison of kappa rate of LS-SVM, (K-Mean^{++}+LS-SVM), (ABC+SVM) and proposed framework

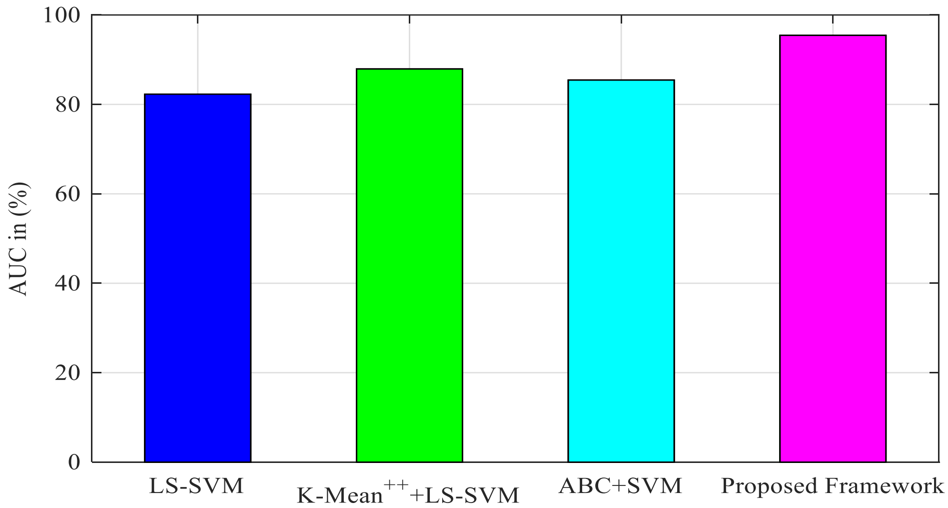


Figure 5.7: Comparison of AUC rate of LS-SVM, (K-Mean^{++}+LS-SVM), (ABC+SVM) and Proposed Framework

Figures 5.5-5.7 demonstrates the results of LS-SVM, (K-Mean^{++}+LS-SVM), (ABC+SVM) and proposed diabetes framework in terms of specificity, kappa and AUC rates. These parameters are also important to investigate the performance of diagnostic system. Figure 5.5 shows the specificity results of proposed framework and other techniques. Its stated that proposed framework achieves higher specificity rate as compared than other techniques. Specificity determines a person with negative test and more importance in diagnostic filed. Hence, its stated that proposed framework is more capable to determine the person with negative diabetes report. Kappa parameter explores the relationship (sensitivity and specificity) rather than statistical significance and kappa value more than 8 should be good. Figure 5.6 illustrates the kappa parameter results of proposed framework, LS-SVM, (K-Mean^{++}+LS-SVM), and (ABC+SVM) techniques. The proposed framework obtains higher kappa rate than others. The results of AUC parameter are demonstrated in Figure 5.7. This parameter shows the trade-off among true positive rate and false positive rate. Its observed that AUC more than 0.9 on the scale of 10 can be consider as outstanding. It is said that proposed frame work achieves higher AUC rate than other technique and proposed framework achieves more than 90% AUC rate. Hence, it is concluded that proposed framework is an effective diabetes prediction system.

The simulation results of proposed diabetes framework are also compared with the state of art techniques/ methods reported in literature. Table 5.7 presents the simulation results of proposed framework and other existing studies using accuracy rate. Total 31 studies are considered for comparing the results of proposed framework. It is analysed that proposed framework achieves 96.57% accuracy rate. The proposed framework improves the accuracy rate of diabetes an average of 5% as compared to nearest one. Several SVM variants are also chosen for performance comparison and it is seen that proposed framework obtains far better accuracy rate. Hence, it is concluded that data imputation and outlier detection techniques having significant role for improving the results of LS-SVM classifier. It is stated that the accuracy rate of LS-SVM classifier increases up to 13% than SVM variants.

Table 5.7: Results of proposed hybrid diabetes framework and other existing studies in literature.

Sr. No.	Study	Algorithm	Accuracy (%)
1	Deng and Kasabov [106]	ESOM	78.4
2	Polat et al. [107]	LS-SVM,	78.21
3	Temurtas et al. [108]	MLNN with LM	79.62
4	Kayaer and Yıldırım [109]	GRNN	80.21
5	Carpenter and Markuzon [110]	ARTMAP-IC	81
6	Temurtas et al. [108]	MLNN with LM	82.37

7	Dogantekin et al. [111]	LDA-ANFIS	84.61
8	Bozkurt et al. [112]	DTDN	76
9	Yilmaz et al. [113]	Modified K-Means Clustering	93.71
11	Dogantekin et al. [111]	Various methods	59.5 and 77.7
12	Ramezani et al. [114]	LANFIS	88.05
13	Orkcu and Bal [115]	Real-coded Genetic Algorithm	77.6
14	Luukka [116]	Similarity Classifier + Feature Extraction	75.97
15	Isa et al. [117]	Clustered-Hybrid MLP	80.59
16	Ozcift and Gulten [118]	Rotation Forest Ensemble Classifier	74.47
17	Aslam et al. [119]	Genetic Programming+K-Nearest Neighbour	80.5
18	Seera and Lim [120]	Fuzzy Max-Min NN-CART Random Forest	78.39
19	Belle et al. [121]	Radial Basis Function Classifier	76.7
20	Wang et al. [122]	Improved Electro magnetism like Mechanism	77.21
21	Seera et al. [133]	Hybrid Fuzzy ARTMAP-CART model	87.64
22	Zhu et al. [85]	Multiple Factors Weighted Combination	93
23	Ding et al. [124]	Extreme Learning Machine	77.63
24	Mohapatra et al. [125]	Improved Cuckoo Search based ELM	78.5
25	Feng et al. [126]	Variable Coded Hierarchical Fuzzy Classification	79.17
26	Luukka [127]	Similarity Classifier using PCA and Entropy	75.82
27	Polat and Gunes [128]	Fuzzy-Artificial immune recognition system	84.42
28	Polat and Gunes [129]	PCA + ANFIS	89.47
29	Ghazavi and Liao [130]	Fuzzy Modelling with Selected Features	77.65
30	Polat et al. [107]	Generalized discriminant analysis Least square SVM	82.05
32	**Our Study**	**(K-Mean^{++} +ABC+LS-SVM)**	**96.57**

5.7 Summary

This chapter presents a hybrid diabetes prediction framework for accurate prediction of diabetes patients. The proposed framework comprises of three different techniques. These techniques are K-Mean++ data imputation, ABC based outlier detection and LS-SVM based classifier. In real world, sometimes attributes of dataset are not containing complete information and also having missing values. The missing values and incomplete information can lead to several other problems such as lack of efficacy, difficulty for managing the data and data analysis. The missing values can be imputed using K-Mean++ based data imputation technique. It is noted that presence of outliers can affect the predictive accuracy. Hence, an ABC based outlier detection technique is developed for determining outlier in this work. These techniques can be applied for transforming the pre-processing data into processed data. Furthermore, LS-SVM classifier is adopted for extracting the pattern from processed data. The performance of proposed framework is evaluated using pima Indian diabetes dataset and this dataset contains 763 missing values and presence of outliers. These problems are handled through K-Mean++ data imputation and ABC based outlier detection techniques. The diabetes prediction is done through LS-SVM classifiers. Several well-known existing studies are considered for comparing the results of proposed framework. Its revealed that proposed framework achieves higher results for diabetes prediction in terms of accuracy, sensitivity, specificity, kappa and AUC parameter.

CHAPTER 6

RULE BASED MONITORING SYSTEM FOR DIABETES

Personal Healthcare Record of patients are widely adopted for diagnosis of different diseases in literature. Furthermore, lack of studies is reported on the usability of PHR in context of diabetes diagnosis. This chapter explores the efficacy of PHR for accurate diagnosis of diabetes. So, this work presents a monitoring system using PHR of diabetes patients. Further, some rules are developed for correctly diagnosis of diabetes. These rules are devised on the basis of PHR information collected through a diabetes application. This application is used for collecting the PHR information from different users also deploying in mobile environment. Hence, the aim of this chapter is to develop a rule-based monitoring system for monitoring and accurate prediction of diabetes disease.

6.1 Chapter Contribution

This chapter develops a monitoring and prediction system for diabetes disease for determining the diabetes risk factor by considering the PHR information. A mobile application is developed for collecting the PHR information and also tracking the users. This app also capable for sending the message to registered users regarding the diabetes prevention and corrective action and further, also facilitates the user for consulting the doctors. The aim of proposed system is to identify the potential risk for diabetes disease and awareness among people. The main contribution of chapter is listed as

1. Developed a diabetes application for collecting the PHR information and integrates into mobile environment.

2. Design a rule base using the PHR information of users.
3. Developed a monitoring and prediction system for accurate diagnosis of diabetes disease.

6.2 Diabetes Monitoring and Prediction System

The proposed diabetes monitoring and prediction system is discussed in this section. The proposed system comprises of three phases. Phase 1 corresponds for collecting the PHR data. Phase 2 responsible for rules-based classification of PHR information. Phase 3 evaluates the performance of proposed rule-based diabetes monitoring and prediction systems. The details of these phases are discussed as below.

6.2.1 Collecting PHR Information

The first phase of the proposed system is to collect the PHR information of different people for accurate diagnosis of diabetes disease. The PHR information is collected through an online diabetes application. Prior to collect the information, several important attributes regarding the diabetes disease are extracted for differentiate diabetes and non-diabetes patients. The details of these attributes are mentioned in Table 6.1. These attributes are BMI, BP, CL, HD, Lifestyle behavior, Prenatal DB and Age. Some other attributes are also designed to classify the diabetes as Type-1 and Type-2 such as excessive thrust, level of urination, felling tired, weight loss, itchiness, blurring vision, eating habit, being unwell, etc. The diabetes app is designed using android platform and integrated into mobile environment. The diabetes app is downloaded on mobile phone and details can be submitted online through this app. Another usability of this app is connecting with the people all time and also taking the feedback and suggestions for diabetes prevention and risk factors from the users. Figure 6.1-6.91illustrates the snaps of the proposed diabetes app.

Table 6.1: Standards for pre diabetic screening

Sr. No.	Standards	Range
1	BMI	> 25
2	BP	> 120/80
3	CL	High/Low
4	HD	Yes/No
5	Lifestyle behavior	Yes/No
6	Prenatal DB	Yes/No
7	Age	> 40

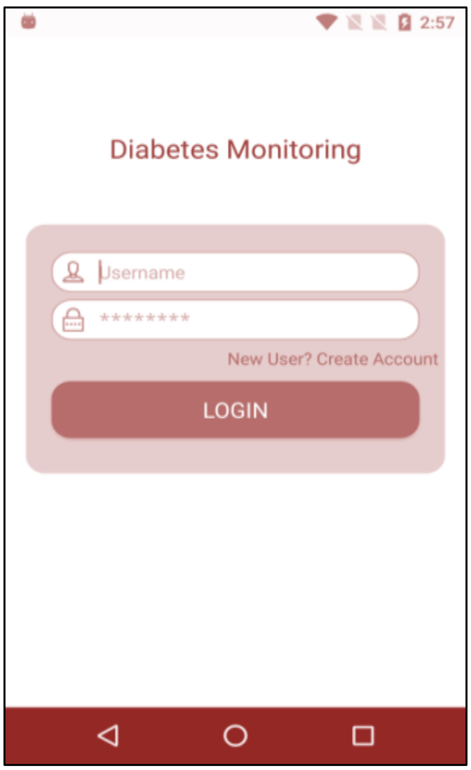


Figure 6.1: LOGIN Details

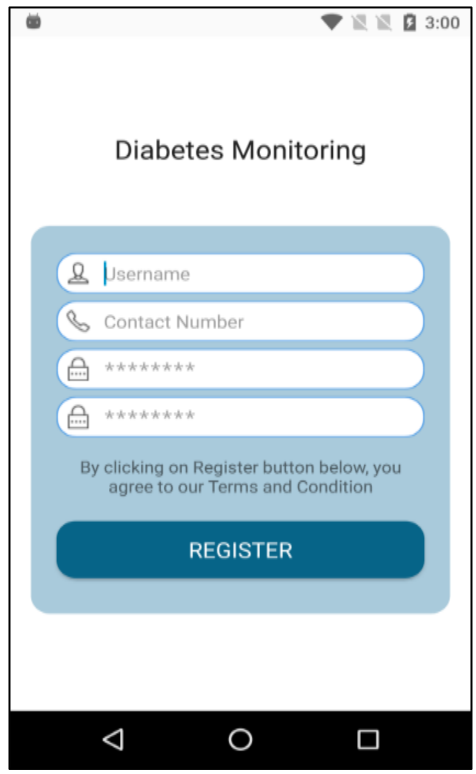


Figure 6.2: User Registration

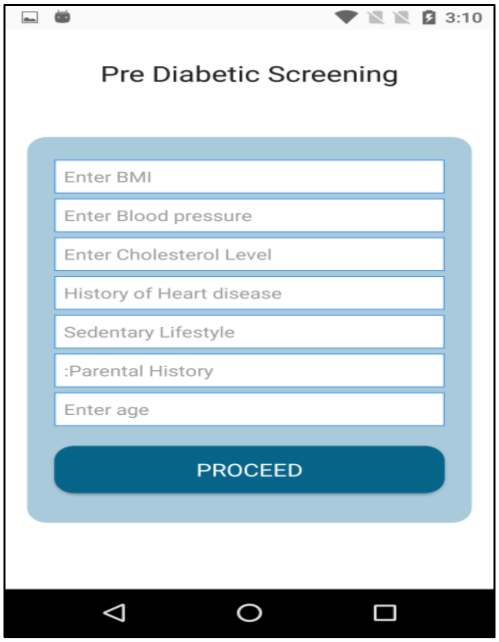


Figure 6.3: Pre-Diabetic Interface

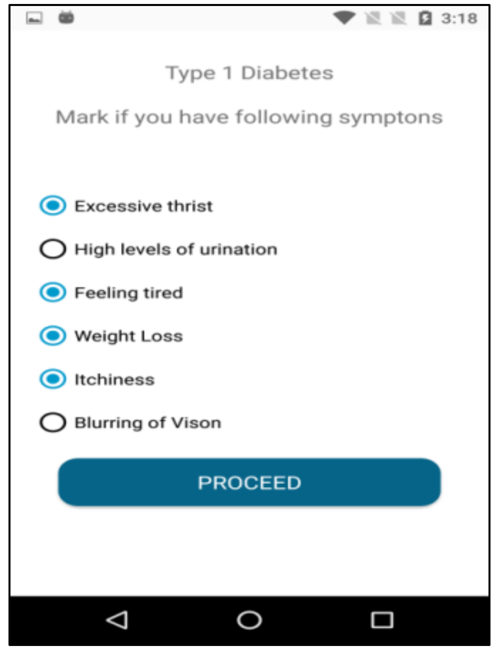


Fig. 6.4: Type- 1 symptoms non clinical

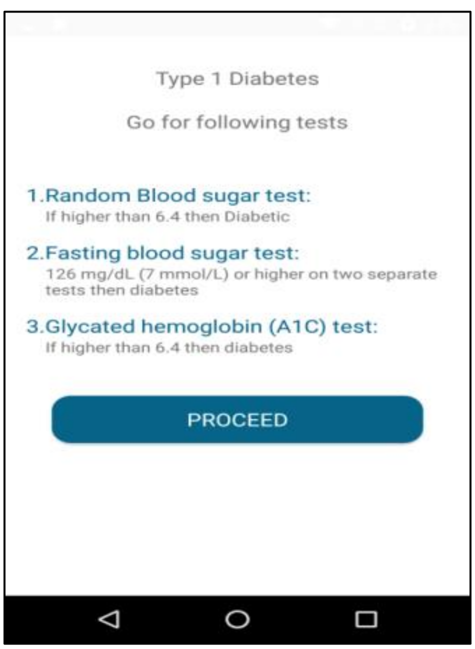


Figure 6.5: Type-1 Clinical symptoms

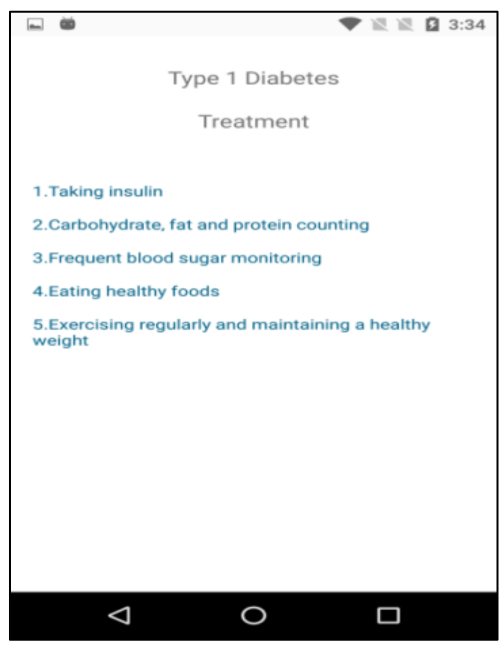
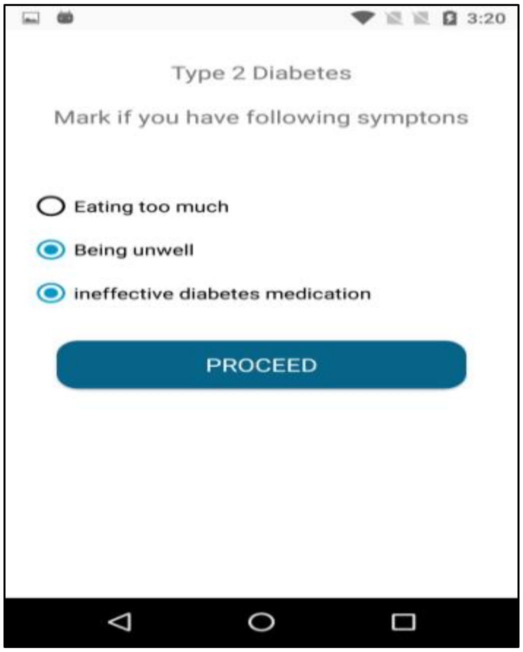


Figure 6.6: Suggestions for Type- 1 diabetes



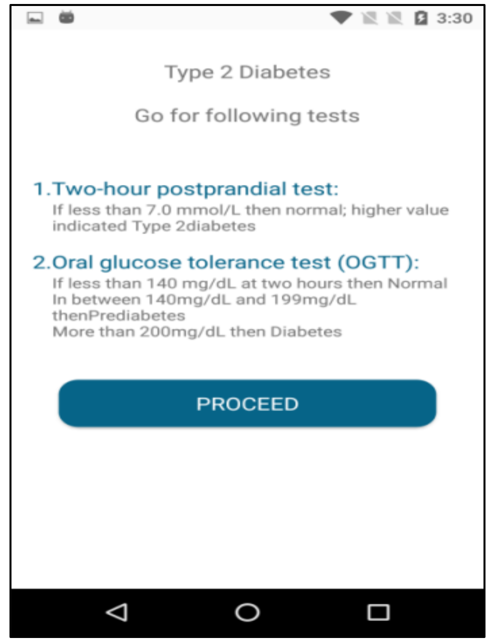


Figure 6.7: Type- 2 diabetes Symptoms

Figure 6.8: Type- 2 diabetes Clinical symptoms

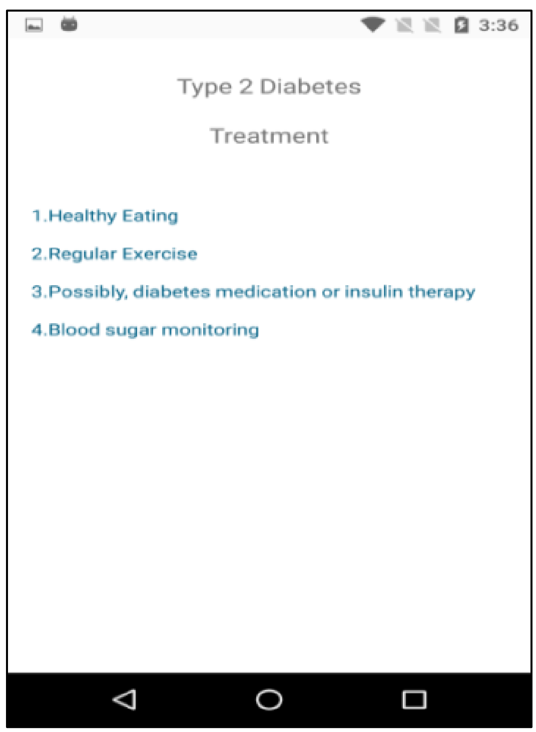


Figure 6.9: Suggestions for Type- 2 diabetes

6.2.2 Rule Base for Diabetes Identification

The phase 2 responsible for designing the rule base for diabetes monitoring and prediction system. These rules are devised using the collected PHR information in phase 1. In this work, twelve rules are developed for diabetes prediction. These rules are as follows.

Rule 1: If ((BMI>25) && (Blood Pressure==Yes) && (CL==high) && (HD==Yes) && (Prenatal DB==Yes) && (lifestyle behavior==Yes) && (Age >40)); Pre-diabetic

Rule 2: If ((BMI>25) && (BP==Yes) && (CL==high) && (History HD==No) && (Prenatal DB==Yes) && (Lifestyle behavior ==Yes) && (Age<40)); Pre-diabetic

Rule 3: If ((BMI<25) && (BP==Yes) && (CL==high) && (History HD==No) && (Prenatal DB==Yes) && (Lifestyle behavior ==Yes) && (Age<40)); Pre-diabetic

Rule 4: If ((BMI<25) && (BP==Yes) && (CL== lower) && (HD==No) && (Prenatal DB==Yes) && (Lifestyle behavior ==Yes) && (Age<40)); Go for diabetic test

Rule 5: If ((BMI<25) && (BP==NO) && (CL== lower) && (HD==No) && (Prenatal DB==Yes) && (Lifestyle behavior ==Yes) && (Age<40)); Go for diabetic test

Rule 6: If ((BMI<25) && (BP==NO) && (CL== lower) && (HD==No) && (Prenatal DB==No) && (Lifestyle behavior ==No) && (Age>40)); Normal

Rule 7: If ((BMI<25) && (BP==NO) && (CL== lower) && (HD==No) && (Prenatal DB==No) && (Sedentary lifestyle==No) && (Age<40)); Normal

Rule 8: If ((LDL>100) && (HDL<40) && (Triglycerides>150)); High Cholesterol

```
Rule 9: If ((LDL>100) || (HDL<40) || (Triglycerides>150)); High
        Cholesterol
Rule 10: If ((LDL<100) && (HDL>40) && (Triglycerides<150)); Low
         Cholesterol
Rule 11: If ((Random test >100) && (Fasting test>126) && (Glycated
         hemoglobin>140)); Type-1 Diabetes
Rule 12: If ((Two-hour test >180) && (Oral glucose tolerance test
         >200)); Type-2 Diabetes
```

Further, it is stated that the afore-mentioned diabetes rules consider the diagnosis measures/standards of diabetes disease. The information regarding the diagnosis measures/standards are presented into Tables 6.2-6.5. Moreover, a panel of doctors is also consulted for certifying the above-mentioned diabetes rules and this panel recommends the rules for diabetes diagnosis.

Table 6.2: Illustrate Cholesterol standards

Standard	Classification	Range
CL	LDL	> 100 mg/dl
	HDL	< 40 mg/dl
	Triglycerides	> 150 mg/dl

Table 6.3: Blood Pressure standards

Standard	Classification	Range
BP	Healthy BP	120/80 mmHg
	Early High BP	In range of 120/80-140/90 mmHg
	High BP	> 140/90 mmHg

Table 6.4: Standards for Type-1 diabetes

Diabetes	Tests	Range
Type-1	Random test	> 100mg/dl
	Fasting test	> 126 mg/dl
	Glycated hemoglobin	> 140 mg/dl

Table 6.5: Standards for Type-2 diabetes

Diabetes Classification	Tests	Range
Type-2	Two-hour test	> 180 mg/dl
	Glucose test	> 200 mg/dl

6.2.3 Performance Evaluation

Phase 3 responsible for evaluating the performance of proposed system. It can be described in terms of diabetes and non-diabetes. Further, this phase also responsible for communication between doctors and patients regarding consultation. The proposed system stores the information in centralized server and this information can be accessed through doctors. The patients' medical report can be checked online and doctor also consults to patients through online mode and sends prescription to patients online and through message.

6.3 Pseudo Code of Diabetes Monitoring and Prediction System

This section illustrates the pseudo code of proposed diabetes monitoring and prediction system and it is given below. Figure 6.10 illustrates the proposed system.

Algorithm 1: Pseudo Code Diabetes Monitoring and Prediction System

1. If user is already registered with decision system, then user click on login option, otherwise click on registration link.
2. After successful login, User enter their personal health details like BMI, blood pressure, cholesterol, history of diabetes etc. to detect pre-diabetic or normal condition
3. If user is pre-diabetic, then it is suggested for predetermined laboratory tests to identify Type-1 or

Type-2 diabetes, otherwise, suggested for workout and control BMI, etc.

4. For Type-1 diabetes, some non-clinical parameters like thirst level, urination level, Loss in weight, tiredness and vision blurring are described. If existence of these parameters is found then go for clinical tests for confirmation of Type-1 diabetes.
5. If clinical tests, confirm Type-1 diabetes, then patients take recommended dose of insulin and also go for Type-2 diabetes.
6. In Type-2 diabetes, go for post prandial test and OGTT, if test positive, patients is affected with Type-2 diabetes and precaution can be taken.
7. If any user needs doctor consultation, then go for doctor consultation link.

6.4 Experimental Section and Results Discussion

This section discusses the experimental results of proposed system. The efficacy of proposed diabetes system is evaluated using two experiments, first experiment is performed on the data collected through diabetes application, while second experiment considers the benchmark pima Indian diabetes dataset. The diabetes application collects the PHR data of 240 people in which 150 people suffers with diabetes and 90 are non-diabetes. The attributes of these datasets are discussed in subsection 6.2.1. The popular machine learning algorithm such as ANN, DT, NB, SVM, LR, RBF-SVM are considered for comparing the simulation results of proposed system. Further, accuracy, sensitivity, specificity and error rate measures are selected as performance parameters for evaluating the simulation results. Initially, the simulation results of aforementioned techniques and proposed diabetes system are represented as confusion matrix. The performance parameters are derived through confusion matrix.

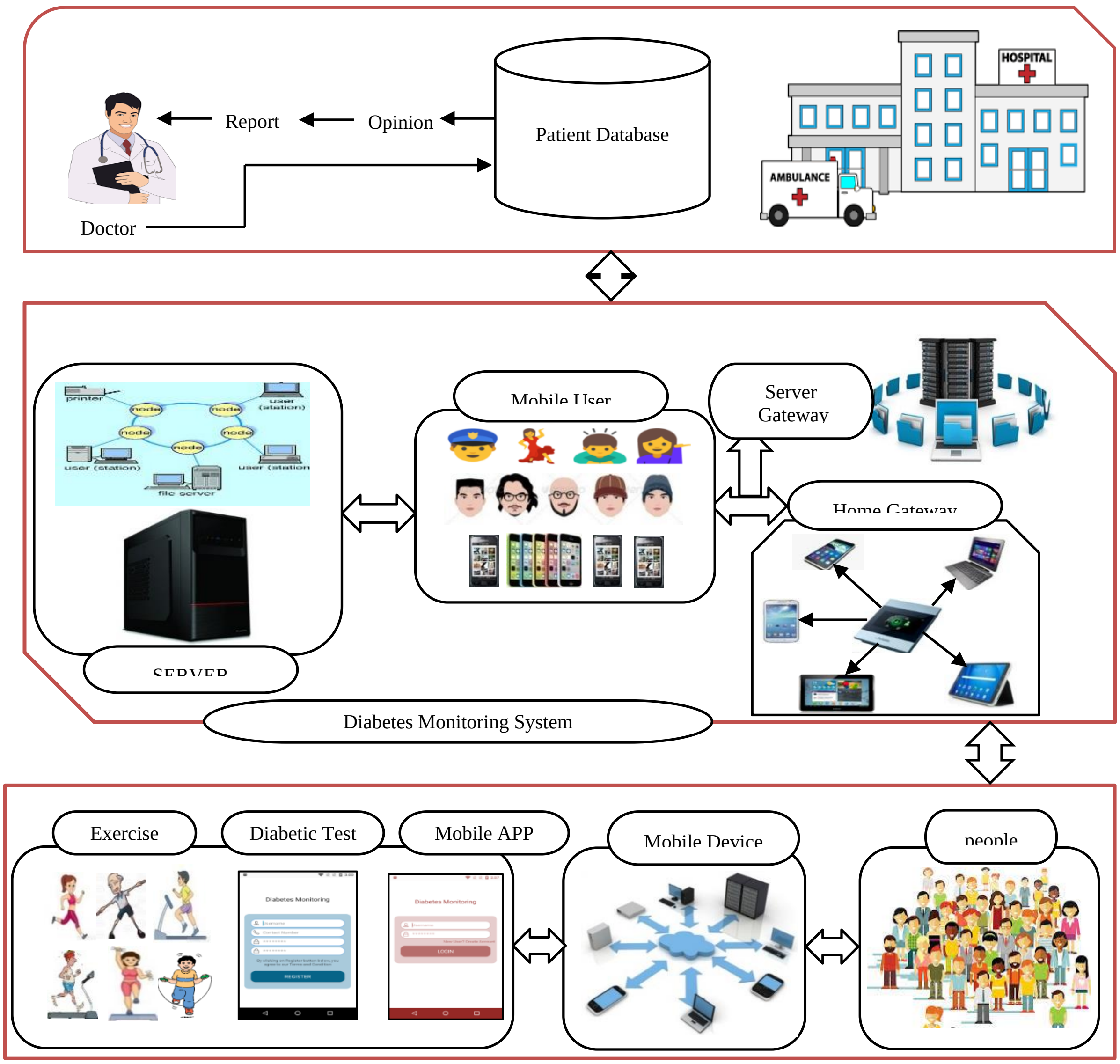


Figure 6.10: Diabetes Monitoring System

Tables 6.6-6.8 present the confusion matrix of ANN, DT, NB, SVM, LR, RBF-SVM and proposed diabetes system. On analyzing the confusion matrix's, it is seen that proposed diabetes system determine the diabetes patient more accurately as compared to ANN, DT, NB, SVM, LR, RBF-

SVM techniques. Whereas, less accurate results are produced by decision tree technique among all other techniques including proposed diabetes monitoring and prediction system.

Table 6.6: Confusion matrix of ANN, DT and NB techniques using collected PHR information

Confusion Matrix		ANN		DT		NB	
		Predicted					
		P	N	P	N	P	N
Actual	P	116	34	113	37	112	28
	N	18	72	20	21	21	69

Table 6.7: Confusion matrix of SVM, LR and RBF-SVM techniques using collected PHR information

Confusion Matrix		SVM		LR		RBF-SVM	
		Predicted					
		P	N	P	N	P	N
Actual	P	124	26	125	25	137	21
	N	18	72	16	74	14	80

Table 6.8: Confusion matrix of proposed diabetes monitoring and prediction system using collected PHR information

Confusion Matrix		Proposed System	
		Predicted	
		Positive	Negative
Actual	Positive	137	13
	Negative	10	80

The comparative analysis of simulation results of proposed diabetes monitoring system and other machine learning techniques is presented in Table 6.9. The proposed diabetes monitoring and prediction system archives more accurate results than other machine learning techniques. The rule base can also enhance the accuracy of proposed system in significant manner. The specificity and sensitivity parameters also confirm the effectiveness of the proposed diabetes system. The proposed system achieves higher sensitivity and specificity rates as compared to rest of techniques. It also analyzed that DT technique provides less significant classification results as compared to other techniques. Furthermore, Figure 6.11 shows the performances of proposed diabetes system and other ML techniques in graphical manner. Hence, it is stated that proposed diabetes monitoring and prediction system obtains significantly better results for diabetes prediction and identification as compared other techniques.

Table 6.9: Comparative analysis of simulation results of proposed system and other machine learning techniques.

Parameters	Techniques						
	ANN	DT	NB	SVM	LR	RBF SVM	Proposed System
Accuracy	78.33	76.25	79.5	81.6	82.9	85.83	90.41
Rank	6	7	5	4	3	2	1
Sensitivity	86.56	84.96	84.21	87.32	88.65	90.72	93.19
Rank	5	7	3	6	4	2	1
Specificity	69.23	65.42	71.13	73.46	74.74	78.35	86.02
Rank	6	7	5	4	3	2	1
Error Rate	21.66	23.75	20.5	18.4	17.1	14.17	9.59
Rank	6	7	5	4	3	2	1
Overall Rank	5.75	7	4.5	4.5	3.25	2	1

The pima Indian diabetes dataset is also considered for evaluating the efficiency of proposed diabetes monitoring and prediction system. The simulation results of proposed diabetes system and other ML techniques are reported in Table 6.10. Simulation results showed that proposed diabetes system achieves first rank among all other techniques. Further, the proposed diabetes monitoring and prediction system significantly improves the accuracy, sensitivity and specificity rates. In turn obtains least error rate than other techniques. It is also analyzed that ANN techniques obtains less significant results for pima Indian diabetes dataset as compared to rest of techniques. Hence, it is stated that proposed diabetes monitoring and prediction system can determine the diabetes disease more efficiently. Figure 6.12 demonstrates the performance of proposed diabetes system and other ML techniques in graphical manner and it is observed that proposed diabetes system obtains good classification results in terms of accuracy, sensitivity, specificity and error rate parameters. It is clearly stated that proposed system attains better results than compared algorithm and determines risk of diabetes in people more accurately.

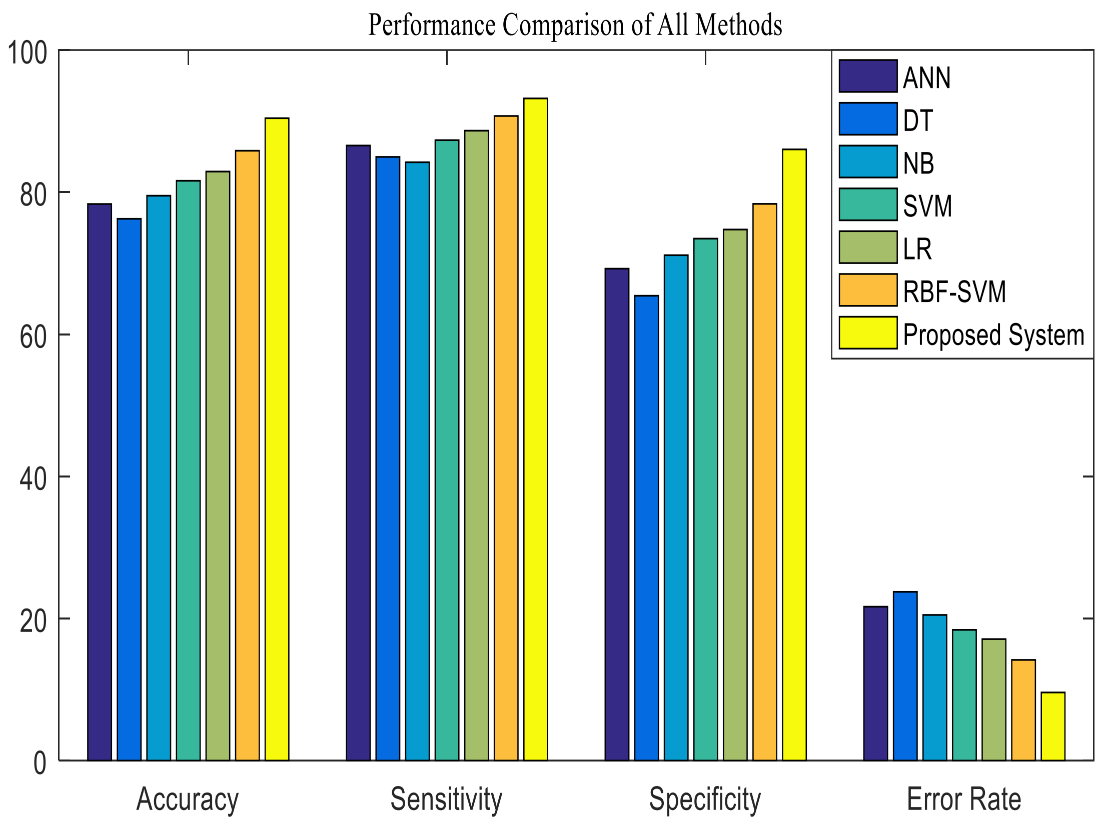


Fig. 6.11: Performance comparison of proposed system and other methods using different performance measures

Table 6.10: Comparative performance analysis of proposed diabetes system and ML techniques for Pima Indian diabetes dataset

Parameters	Techniques						
	ANN	DT	NB	SVM	LR	RBF SVM	Proposed System
Accuracy	65.51	62.45	68.19	72.06	62.83	76.56	85.23
Rank	5	7	4	3	6	2	1
Sensitivity	65.23	85.14	84.21	77.02	65.38	86.14	90.16
Rank	7	3	4	5	6	2	1
Specificity	64.48	65.56	71.13	74.56	72.84	81.64	87.34
Rank	7	6	5	3	4	2	1
Error Rate	34.59	37.55	31.81	27.94	37.17	23.44	14.77
Rank	5	7	4	3	6	2	1
Overall Rank	6	5.75	4.25	3.5	5.5	2	1

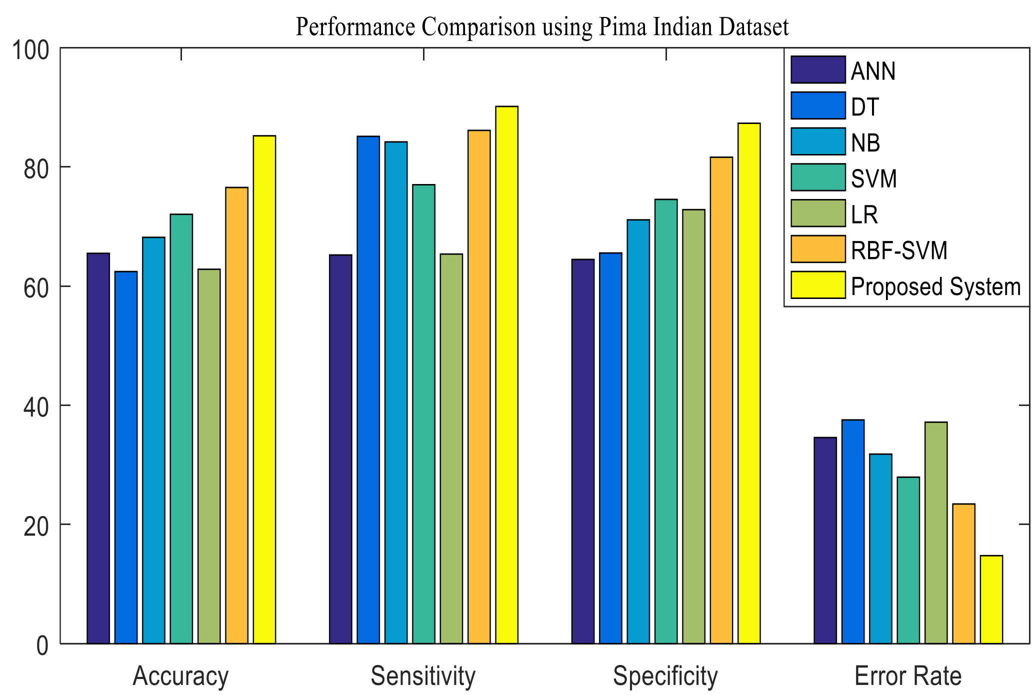


Figure 6.12: Results of proposed diabetes system and ML techniques for Pima Indian diabetes dataset

6.5 Summary

This chapter presents a diabetes monitoring and prediction system for accurate diagnosis of diabetes disease. The proposed system comprises of three phases-data collection, rule base and evaluation. The data collection phases responsible for collection of data. The data is collected through a diabetes app. The rule base phase corresponds for designing the rules to predict the diabetes. In this work, twelve rules are designed for accurate prediction of diabetes on the behalf of collected PHR information of people. The evaluation phase corresponds for evaluating the performance of rule base and consolation of diabetes patients with doctor. The accuracy, sensitivity, specificity and error rate parameters are considered as performance measures for evaluating the performance of proposed diabetes system. It is seen that proposed diabetes monitoring and prediction system achieves good results than ML based techniques. Further, pima Indian diabetes dataset is also considered to examine the efficacy of proposed diabetes system and obtained better diabetes prediction results.

CHAPTER 7

CONCLUSION AND FUTURE SCOPE OF WORK

This thesis presents new classifiers and diagnostic models for accurate prediction of diabetes disease. The aim of this research work is to improve the diagnostic accuracy of prediction models and also determine the optimal features for diabetes disease. The primary objectives of this thesis work are listed as deployment of automatic diagnostic model for the diagnostic and prediction of diabetes disease using real data; developing an efficient machine learning technique to improve the accuracy rate of existing technique using feature extraction weighting method and design the personal health record-based model for diagnosis of diabetes disease using real data. The aforementioned objectives are addresses in this thesis work. To achieve the objectives, four prediction models are developed in this work. These prediction models are KHM-AW-SVM diagnostic system, ABC-DNN based diagnostic model, Hybrid diabetes disease prediction framework (K-Mean^{++}+ABC+LS-SVM) and rule-based monitoring system for diabetes prediction and diagnosis. The entire thesis work is divided into four chapters.

The chapter three considers the accuracy issue of SVM classifier in context of diagnostic system. This issue of SVM is addressed through selection of optimal attributes for diagnosis process. So, in this work, KHM-AW method is developed for computing the weight of attributes and further, to determine the non-relevant attributes. The KHM method is used for measuring the label of data object. Further, a weight function is designed for computing the weight of attributes through weight coefficients and cluster centers. This method is integrated into SVM classifier for accurate prediction of diabetes diseases. The results are evaluated using 10-fold and 50-50% training-testing methods. Both all attributes and optimum attributes of diabetes are considered for evaluating the performance of proposed KHM-AW-SVM model. Results confirmed that attribute selection method provides advantage over SVM classifier. Furthermore, it is conferred that 10-fold method obtains better results than 50-50% training-testing method using all attributes and optimum attributes.

ABC-DNN based diagnostic model for the diagnosis and prediction of diabetes disease is proposed in chapter four. The proposed diagnostic model is designed using ABC based feature selection method and deep neural network technique. In proposed model, ABC based feature selection method is used to determine the relevant features of diabetes disease. Further, the DNN technique is adopted for diagnosis and prediction of diabetes disease. The performance of proposed diagnostic model is evaluated using Pima Indian Diabetes dataset. The different performance measures like accuracy, sensitivity, specificity, AUC, F-measure and Kappa are considered to assess the performance of ABC-DNN based diagnostic model. Furthermore, experimental results are evaluated using 10-fold and 50-50% training-testing techniques. It is observed that ABC-DNN based diagnostic model provides better results than DNN method. The experimental results of ABC-DNN model is also compared with thirty-one existing diabetes studies. It also revealed that proposed diagnostic model achieves higher accuracy rate than existing studies. Furthermore, it is noticed that 10-fold method is more suitable than 50-50% training-testing method.

The chapter five presents a hybrid diabetes prediction framework for accurate prediction of diabetes patients. The proposed framework comprises of three different techniques. These techniques are K-Mean++ data imputation, ABC based outlier detection and LS-SVM based classifier. In real world, sometimes attributes of dataset are not containing complete information and also having missing values. The missing values and incomplete information can lead to several other problems such as lack of efficacy, difficulty for managing the data and data analysis. The missing values can be imputed using K-Mean++ based data imputation technique. It is noted that presence of outliers can affect the predictive accuracy. Hence, an ABC based outlier detection technique is developed for determining outlier in this work. These techniques can be applied for transforming the pre-processing data into processed data. Furthermore, LS-SVM classifier is adopted for extracting the pattern from processed data. The performance of proposed framework is evaluated using pima Indian diabetes dataset and this dataset contains 763 missing values and presence of outliers. These problems are handled through K-Mean++ data imputation and ABC based outlier detection techniques. The diabetes prediction is done through LS-SVM classifiers. Several well-known existing studies are considered for comparing the results of proposed framework. Its revealed that proposed framework achieves higher results for diabetes prediction in terms of accuracy, sensitivity, specificity, kappa and AUC parameter.

The chapter six presents a diabetes monitoring and prediction system for accurate diagnosis of diabetes disease. The proposed system comprises of three phases-data collection, rule base and evaluation. The data collection phases responsible for collection of data. The data is collected through a diabetes app. The rule base phase corresponds for designing the rules to predict the diabetes. In this work, twelve rules are designed for accurate prediction of diabetes on the behalf of collected PHR information of people. The evaluation phase corresponds for evaluating the performance of rule base and consolation of diabetes patients with doctor. The accuracy, sensitivity, specificity and error rate parameters are considered as performance measures for evaluating the performance of proposed diabetes system. It is seen that proposed diabetes monitoring and prediction system achieves good results than ML based techniques. Further, pima Indian diabetes dataset is also considered to examine the efficacy of proposed diabetes system and obtained better diabetes prediction results.

In this thesis, four prediction models are developed for diagnosis and prediction of diabetes disease. The feature selection issues are also addressed through KHM and ABC based feature selection algorithms. The aim of these algorithms is to determine the optimal features for prediction of diabetes. Further, a data imputation method based on K-Mean++ is also developed for handling the missing value imputations. The results are evaluated using 50-50 percent training-test and 10-cross fold method. It is noticed that K-Mean^{++}+ABC+LS-SVM outperforms than other compared method.

7.1 Future Scope

In future, some other diseases like thyroid, cancer, stroke, etc., will be considerd for diagnosis and prediction task and will be developed more suitable and efficient prediction system. The feature selection and outlier deetction are major problems associated with medical datasets. These problem will be considered in the future work and will provide efficient solution for these problems. Furthermore, oversampled and undersampled datasets will be considered for the diagnosis task.

REFERENCES

1. Doi, K. (2007). Computer-aided diagnosis in medical imaging: historical review, current status and future potential. Computerized medical imaging and graphics, 31(4-5), 198-211.
2. Rangayyan, R. M., Ayres, F. J., & Desautels, J. L. (2007). A review of computer-aided diagnosis of breast cancer: Toward the detection of subtle signs. Journal of the Franklin Institute, 344(3-4), 312-348.
3. Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C., & Hurdle, J. F. (2008). Extracting information from textual documents in the electronic health record: a review of recent research. Yearbook of medical informatics, 17(01), 128-144.
4. Nilashi, M., Bin Ibrahim, O., Ahmadi, H., & Shahmoradi, L. (2017). An analytical method for diseases prediction using machine learning techniques. Computers & Chemical Engineering, 106, 212-223.
5. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal, 15, 104-116.
6. Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2017). Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE journal of biomedical and health informatics, 22(5), 1589-1604.
7. Nahar, J., Imam, T., Tickle, K. S., & Chen, Y. P. P. (2013). Computational intelligence for heart disease diagnosis: A medical knowledge driven approach. Expert Systems with Applications, 40(1), 96-104
8. World Health Organization. http://www.who.int
9. Alice, S., & Balachandran, S. (2015), Performance Analysis of Training Algorithms of Multilayer Perceptrons in Diabetes Prediction. Proceedings of the International Conference on Advances in Computer Engineering and Applications (ICACEA), Ghaziabad, India.
10. Zheng, J. W., Zhang, Z. B., Wu, T. H., & Zhang, Y. (2007). A wearable mobihealth care system supporting real-time diagnosis and alarm. Medical Biological Engineering Computing, 45(9), 877–885.

11. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., and Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Comput. Struct. Biotechnol. J. 15, 104–116. doi: 10.1016/j.csbj.2016.12.005
12. Yadav, G., Kumar, Y. and Sahoo, G., (2012), Predication of Parkinson's disease using data mining methods: A comparative analysis of tree, statistical and support vector machine classifiers. In Computing and Communication Systems (NCCCS), 2012 National Conference on (pp. 1-8). IEEE.
13. Kumar, Y., & Sahoo, G. (2013). Prediction of different types of liver diseases using rule based classification model. Technology and Health Care, 21(5), 417-432.
14. Yue, C., Xin, L., Kewen, X., and Chang, S. (2008). “An intelligent diagnosis to type 2 diabetes based on QPSO algorithm and WLS-SVM,” in Proceedings of the 2008 IEEE International Symposium on Intelligent Information Technology Application Workshops, Washington, DC. doi: 10.1109/IITA.Workshops.2008.36
15. Duygu, ç., and Esin, D. (2011). An automatic diabetes diagnosis system based on LDA-wavelet support vector machine classifier. Expert Syst. Appl. 38, 8311–8315.
16. Sahoo, A. J., & Kumar, Y. (2014). Seminal quality prediction using data mining methods. Technology and Health Care, 22(4), 531-545.
17. Gambhir, S., Malik, S. K., & Kumar, Y. (2016). Role of soft computing approaches in healthcare domain: a mini review. Journal of medical systems, 40(12), 287.
18. Ozcift, A., and Gulten, A. (2011). Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms. Comput. Methods Programs Biomed. 104, 443–451. doi: 10.1016/j.cmpb.2011.03.018.
19. Gambhir, S., Malik, S. K., & Kumar, Y. (2017). PSO-ANN based diagnostic model for the early detection of dengue disease. New Horizons in Translational Medicine, 4(1-4), 1-8.
20. Kumar, Y., Yadav, G., Singh, P. K., & Arora, P. (2019). A PHR-Based System for Monitoring Diabetes in Mobile Environment. In Mobile Solutions and Their Usefulness in Everyday Life (pp. 129-144). Springer, Cham.
21. Gambhir, S., Malik, S. K., & Kumar, Y. (2018). The Diagnosis of Dengue Disease: An Evaluation of Three Machine Learning Approaches. International Journal of Healthcare Information Systems and Informatics (IJHISI), 13(3), 1-19.

22. Lin, Y., Qin, H., Chen, R., Liu, Q., Liu, H., & Dong, S. (2019). A comprehensive clinical diagnostic score system for prediction of coronary artery spasm in patients with acute chest pain. IJC Heart & Vasculature, 22, 205-209.
23. Sabra, S., Malik, K. M., Afzal, M., Sabeeh, V., & Charaf Eddine, A. (2020). A hybrid knowledge and ensemble classification approach for prediction of venous thromboembolism. Expert Systems, 37(1), e12388.
24. Acharjya, D. P. (2020). A Hybrid Scheme for Heart Disease Diagnosis Using Rough Set and Cuckoo Search Technique. Journal of Medical Systems, 44(1), 27.
25. Abdar, M., Książek, W., Acharya, U. R., Tan, R. S., Makarenkov, V., & Pławiak, P. (2019). A new machine learning technique for an accurate diagnosis of coronary artery disease. Computer methods and programs in biomedicine, 179, 104992.
26. Narayan, S., & Sathiyamoorthy, E. (2019). A novel recommender system based on FFT with machine learning for predicting and identifying heart diseases. Neural Computing and Applications, 31(1), 93-102.
27. Bucholc, M., Ding, X., Wang, H., Glass, D. H., Wang, H., Prasad, G., ... & Finn, D. P. (2019). A practical computerized decision support system for predicting the severity of Alzheimer's disease of an individual. Expert systems with applications, 130, 157-171.
28. Nilashi, M., Ahmadi, H., Shahmoradi, L., Ibrahim, O., & Akbari, E. (2019). A predictive method for hepatitis disease diagnosis using ensembles of neuro-fuzzy technique. Journal of infection and public health, 12(1), 13-20.
29. Zhao, D., Liu, H., Zheng, Y., He, Y., Lu, D., & Lyu, C. (2019). A reliable method for colorectal cancer prediction based on feature selection and support vector machine. Medical & biological engineering & computing, 57(4), 901-912.
30. Nilashi, M., Ibrahim, O., Samad, S., Ahmadi, H., Shahmoradi, L., & Akbari, E. (2019). An analytical method for measuring the Parkinson’s disease progression: A case on a Parkinson’s telemonitoring dataset. Measurement, 136, 545-557.
31. Ali, L., Rahman, A., Khan, A., Zhou, M., Javeed, A., & Khan, J. A. (2019). An Automated Diagnostic System for Heart Disease Prediction Based on ${\chi^{2}} $ Statistical Model and Optimally Configured Deep Neural Network. IEEE Access, 7, 34938-34945.

32. Kanimozhi, U., Ganapathy, S., Manjula, D., & Kannan, A. (2019). An intelligent risk prediction system for breast cancer using fuzzy temporal rules. National Academy Science Letters, 42(3), 227-232.

33. Gokulnath, C. B., & Shantharajah, S. P. (2019). An optimized feature selection based on genetic approach and support vector machine for heart disease. Cluster Computing, 22(6), 14777-14787.

34. Kirk, R. A., & Kirk, D. A. (2017, July). Introducing a Decision Making Framework to Help Users Detect, Evaluate, Assess, and Recommend (DEAR) Action Within Complex Sociotechnical Environments. In International Conference on Human Interface and the Management of Information (pp. 223-239). Springer, Cham.

35. Devarajan, M., & Ravi, L. (2018). Intelligent cyber-physical system for an efficient detection of Parkinson disease using fog computing. Multimedia Tools and Applications, 1-25.

36. Peker, M. (2016). A decision support system to improve medical diagnosis using a combination of k-medoids clustering based attribute weighting and SVM. Journal of medical systems, 40(5), 116.

37. Huda, S., Yearwood, J., Jelinek, H. F., Hassan, M. M., Fortino, G., & Buckland, M. (2016). A hybrid feature selection with ensemble classification for imbalanced healthcare data: A case study for brain tumor diagnosis. IEEE access, 4, 9145-9154.

38. Aličković, E., & Subasi, A. (2017). Breast cancer diagnosis using GA feature selection and Rotation Forest. Neural Computing and Applications, 28(4), 753-763.

39. Abdar, M., Książek, W., Acharya, U. R., Tan, R. S., Makarenkov, V., & Pławiak, P. (2019). A new machine learning technique for an accurate diagnosis of coronary artery disease. Computer methods and programs in biomedicine, 179, 104992.

40. Yadav, S. S., & Jadhav, S. M. (2020). Detection of common risk factors for diagnosis of cardiac arrhythmia using machine learning algorithm. Expert Systems with Applications, 163, 113807.

41. Masood, A., Sheng, B., Li, P., Hou, X., Wei, X., Qin, J., & Feng, D. (2018). Computer-assisted decision support system in pulmonary cancer detection and stage classification on CT images. Journal of biomedical informatics, 79, 117-128.

42. Yadav, D. C., & Pal, S. (2020). Prediction of thyroid disease using decision tree ensemble method. Human-Intelligent Systems Integration, 1-7.

43. Liu, T., Fan, W., & Wu, C. (2019). A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset. Artificial Intelligence in Medicine, 101, 101723.

44. Fang, G., Liu, W., & Wang, L. (2020). A machine learning approach to select features important to stroke prognosis. Computational Biology and Chemistry, 88, 107316.

45. Chen, Y. C., Suzuki, T., Suzuki, M., Takao, H., Murayama, Y., & Ohwada, H. (2017). Building a classifier of onset stroke prediction using random tree algorithm. International Journal of Machine Learning and Computing, 7(4), 61-66.

46. Govindarajan, P., Soundarapandian, R. K., Gandomi, A. H., Patan, R., Jayaraman, P., & Manikandan, R. (2020). Classification of stroke disease using machine learning algorithms. Neural Computing and Applications, 32(3), 817-828.

47. Arslan, A. K., Colak, C., & Sarihan, M. E. (2016). Different medical data mining approaches based prediction of ischemic stroke. Computer methods and programs in biomedicine, 130, 87-92.

48. Goyal, M. (2018, July). Long short-term memory recurrent neural network for stroke prediction. In International Conference on Machine Learning and Data Mining in Pattern Recognition (pp. 312-323). Springer, Cham.

49. Li, X., Wu, M., Sun, C., Zhao, Z., Wang, F., Zheng, X., ... & Zou, J. (2020). Using machine learning to predict stroke-associated pneumonia in Chinese acute ischaemic stroke patients. European Journal of Neurology, 27(8), 1656-1663.

50. Heo, J., Yoon, J. G., Park, H., Kim, Y. D., Nam, H. S., & Heo, J. H. (2019). Machine learning–based model for prediction of outcomes in acute stroke. Stroke, 50(5), 1263-1265.

51. Liu, Y., Yin, B., & Cong, Y. (2020). The Probability of Ischaemic Stroke Prediction with a Multi-Neural-Network Model. Sensors, 20(17), 4995.

52. Monteiro, M., Fonseca, A. C., Freitas, A. T., e Melo, T. P., Francisco, A. P., Ferro, J. M., & Oliveira, A. L. (2018). Using machine learning to improve the prediction of functional outcome in ischemic stroke patients. IEEE/ACM transactions on computational biology and bioinformatics, 15(6), 1953-1959.

53. Shah, T., Yavari, A., Mitra, K., Saguna, S., Jayaraman, P. P., Rabhi, F., & Ranjan, R. (2016). Remote health care cyber-physical system: quality of service (QoS) challenges and opportunities. IET Cyber-Physical Systems: Theory & Applications, 1(1), 40-48.
54. Nandyala, C. S., & Kim, H. K. (2016). From cloud to Fog and IoT-based real-time U-healthcare monitoring for smart homes and hospitals. International Journal of Smart Home, 10(2), 187-196.
55. Costanzo, A., Faro, A., Giordano, D., & Pino, C. (2016, January). Mobile cyber physical systems for health care: Functions, ambient ontology and e-diagnostics. In Consumer Communications & Networking Conference (CCNC), 2016 13th IEEE Annual (pp. 972-975). IEEE.
56. Oluwagbemi, O., Oluwagbemi, F., & Abimbola, O. (2016). Ebinformatics: Ebola fuzzy informatics systems on the diagnosis, prediction and recommendation of appropriate treatments for Ebola virus disease (EVD). Informatics in Medicine Unlocked, 2, 12-37.
57. Sood, S. K., & Mahajan, I. (2018). Fog-cloud based cyber-physical system for distinguishing, detecting and preventing mosquito borne diseases. Future Generation Computer Systems.
58. Thota, C., Sundarasekar, R., Manogaran, G., Varatharajan, R., & Priyan, M. K. (2018). Centralized fog computing security platform for IoT and cloud in healthcare system. In Fog Computing: Breakthroughs in Research and Practice (pp. 365-378). IGI global.
59. Venčkauskas, A., Morkevicius, N., Jukavičius, V., Damaševičius, R., Toldinas, J., & Grigaliūnas, Š. (2019). An edge-fog secure self-authenticable data transfer protocol. Sensors, 19(16), 3612.
60. Vijayakumar, V., Malathi, D., Subramaniyaswamy, V., Saravanan, P., & Logesh, R. (2019). Fog computing-based intelligent healthcare system for the detection and prevention of mosquito-borne diseases. Computers in Human Behavior, 100, 275-285.
61. Saxena, S. K., Elahi, A., Gadugu, S., & Prasad, A. K. (2016). Zika virus outbreak: an overview of the experimental therapeutics and treatment. Virusdisease, 27(2), 111-115.
62. Ginier, M., Neumayr, A., Günther, S., Schmidt-Chanasit, J., & Blum, J. (2016). Zika without symptoms in returning travellers: What are the implications?. Travel medicine and infectious disease, 14(1), 16-20.

63. Pabbaraju, K., Wong, S., Gill, K., Fonseca, K., Tipples, G. A., & Tellier, R. (2016). Simultaneous detection of Zika, Chikungunya and Dengue viruses by a multiplex real-time RT-PCR assay. Journal of Clinical Virology, 83, 66-71.
64. Campion, M., Bina, C., Pozniak, M., Hanson, T., Vaughan, J., Mehus, J., Hanson, S., Cronquist, L., Feist, M., Ranganathan, P. and Kaabouch, N. (2016, December). Predicting West Nile Virus (WNV) occurrences in North Dakota using data mining techniques. In Future Technologies Conference (FTC) (pp. 310-317). IEEE.
65. Lambert, B., Sikulu-Lord, M. T., Mayagaya, V. S., Devine, G., Dowell, F., & Churcher, T. S. (2018). Monitoring the Age of Mosquito Populations Using Near-Infrared Spectroscopy. Scientific reports, 8(1), 5274.
66. Kirk, R. A., & Kirk, D. A. (2017, July). Introducing a Decision Making Framework to Help Users Detect, Evaluate, Assess, and Recommend (DEAR) Action Within Complex Sociotechnical Environments. In International Conference on Human Interface and the Management of Information (pp. 223-239). Springer, Cham.
67. Devarajan, M., & Ravi, L. (2018). Intelligent cyber-physical system for an efficient detection of Parkinson disease using fog computing. Multimedia Tools and Applications, 1-25.
68. Kaur, P., Kumar, R., & Kumar, M. (2019). A healthcare monitoring system using random forest and internet of things (IoT). Multimedia Tools and Applications, 1-12.
69. Parthasarathy, P., & Vivekanandan, S. (2018). A typical IoT architecture-based regular monitoring of arthritis disease using time wrapping algorithm. International Journal of Computers and Applications, 1-11.
70. Tuli, S., Basumatary, N., Gill, S. S., Kahani, M., Arya, R. C., Wander, G. S., & Buyya, R. (2019). HealthFog: An ensemble deep learning based Smart Healthcare System for Automatic Diagnosis of Heart Diseases in integrated IoT and fog computing environments. Future Generation Computer Systems.
71. Priyadarshini, R., Barik, R., & Dubey, H. (2018). DeepFog: Fog Computing-Based Deep Neural Architecture for Prediction of Stress Types, Diabetes and Hypertension Attacks. Computation, 6(4), 62.

72. Jabeen, F., Maqsood, M., Ghazanfar, M. A., Aadil, F., Khan, S., Khan, M. F., & Mehmood, I. (2019). An IoT based efficient hybrid recommender system for cardiovascular disease. Peer-to-Peer Networking and Applications, 1-14.
73. Sood, S. K., & Mahajan, I. (2018). A Fog Assisted Cyber-Physical Framework for Identifying and Preventing Coronary Heart Disease. Wireless Personal Communications, 101(1), 143-165.
74. Gu, Z., Jiang, Y., Zhou, M., Gu, M., Song, X., & Sha, L. (2018). A Cyber-Physical System Framework for Early Detection of Paroxysmal Diseases. IEEE Access, 6, 34834-34845.
75. Sood, S. K., & Mahajan, I. (2018). IoT-Fog-Based Healthcare Framework to Identify and Control Hypertension Attack. IEEE Internet of Things Journal, 6(2), 1920-1927.
76. Lakshmanaprabu, S. K., Shankar, K., Ilayaraja, M., Nasir, A. W., Vijayakumar, V., & Chilamkurti, N. (2019). Random forest for big data classification in the internet of things using optimal features. International Journal of Machine Learning and Cybernetics, 1-10.
77. Anand, L., & Ibrahim, S. S. (2018). HANN: A Hybrid Model for Liver Syndrome Classification by Feature Assortment Optimization. Journal of medical systems, 42(11), 211.
78. Devarajan, M., & Ravi, L. (2018). Intelligent cyber-physical system for an efficient detection of Parkinson disease using fog computing. Multimedia Tools and Applications, 1-25.
79. Kaur, P., Kumar, R., & Kumar, M. (2019). A healthcare monitoring system using random forest and internet of things (IoT). Multimedia Tools and Applications, 1-12.
80. Jabeen, F., Maqsood, M., Ghazanfar, M. A., Aadil, F., Khan, S., Khan, M. F., & Mehmood, I. (2019). An IoT based efficient hybrid recommender system for cardiovascular disease. Peer-to-Peer Networking and Applications, 1-14.
81. Tuli, S., Basumatary, N., Gill, S. S., Kahani, M., Arya, R. C., Wander, G. S., & Buyya, R. (2019). HealthFog: An ensemble deep learning based Smart Healthcare System for Automatic Diagnosis of Heart Diseases in integrated IoT and fog computing environments. Future Generation Computer Systems.
82. Priyadarshini, R., Barik, R., & Dubey, H. (2018). DeepFog: Fog Computing-Based Deep Neural Architecture for Prediction of Stress Types, Diabetes and Hypertension Attacks. Computation, 6(4), 62.

83. Yuvaraj, N., & SriPreethaa, K. R. (2019). Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster. Cluster Computing, 22(1), 1-9.

84. Xie, J., & Wang, Q. (2020). Benchmarking machine learning algorithms on blood glucose prediction for Type 1 Diabetes in comparison with classical time-series models. IEEE Transactions on Biomedical Engineering.

85. Zhu, J., Xie, Q., & Zheng, K. (2015). An improved early detection method of type-2 diabetes mellitus using multiple classifier system. *Information Sciences*, *292*, 1-14.

86. Devi, R. D. H., Bai, A., & Nagarajan, N. (2020). A novel hybrid approach for diagnosing diabetes mellitus using farthest first and support vector machine algorithms. *Obesity Medicine*, *17*, 100152.

87. Maniruzzaman, M., Rahman, M. J., Ahammed, B., & Abedin, M. M. (2020). Classification and prediction of diabetes disease using machine learning paradigm. *Health Information Science and Systems*, *8*(1), 7.

88. Wang, M., & Chen, H. (2020). Chaotic multi-swarm whale optimizer boosted support vector machine for medical diagnosis. *Applied Soft Computing*, *88,* 105946.

89. Singh, N., & Singh, P. (2020). Stacking-based multi-objective evolutionary ensemble framework for prediction of diabetes mellitus. *Biocybernetics and Biomedical Engineering*, *40*(1), 1-22.

90. Manikandan, K. (2019). Diagnosis of diabetes diseases using optimized fuzzy rule set by grey wolf optimization. *Pattern Recognition Letters*, *125*, 432-438.

91. Artzi, N. S., Shilo, S., Hadar, E., Rossman, H., Barbash-Hazan, S., Ben-Haroush, A., ... & Segal, E. (2020). Prediction of gestational diabetes based on nationwide electronic health records. *Nature Medicine*, *26*(1), 71-76.

92. Bernardini, M., Romeo, L., Misericordia, P., & Frontoni, E. (2019). Discovering the type 2 diabetes in electronic health records using the sparse balanced support vector machine. *IEEE Journal of Biomedical and Health Informatics*, *24*(1), 235-246.

93. Li, K., Daniels, J., Liu, C., Herrero, P., & Georgiou, P. (2019). Convolutional recurrent neural networks for glucose prediction. *IEEE journal of biomedical and health informatics*, *24*(2), 603-613.

94. Aliberti, A., Pupillo, I., Terna, S., Macii, E., Di Cataldo, S., Patti, E., & Acquaviva, A. (2019). A multi-patient data-driven approach to blood glucose prediction. *IEEE Access*, *7*, 69311-69325.

95. Fitriyani, N. L., Syafrudin, M., Alfian, G., & Rhee, J. (2019). Development of disease prediction model based on ensemble learning approach for diabetes and hypertension. *IEEE Access*, *7*, 144777-144789.

96. Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers. *IEEE Access*, *8*, 76516-76531.

97. Chen, L., Lu, D., Zhu, M., Muzammal, M., Samuel, O. W., Huang, G., ... & Wu, H. (2019). OMDP: An ontology-based model for diagnosis and treatment of diabetes patients in remote healthcare systems. *International Journal of Distributed Sensor Networks*, *15*(5), 1550147719847112.

98. Abdel-Basset, M., Manogaran, G., Gamal, A., & Chang, V. (2019). A novel intelligent medical decision support model based on soft computing and IoT. *IEEE Internet of Things Journal*, *7*(5), 4160-4170.

99. Alirezaei, M., Niaki, S. T. A., & Niaki, S. A. A. (2019). A bi-objective hybrid optimization algorithm to reduce noise and data dimension in diabetes diagnosis using support vector machines. *Expert Systems with Applications*, *127*, 47-57.

100. Singh, A., & Gupta, G. (2019). ANT_FDCSM: A novel fuzzy rule miner derived from ant colony meta-heuristic for diagnosis of diabetic patients. *Journal of Intelligent & Fuzzy Systems*, *36*(1), 747-760.

101. Jayashree, J., & Kumar, S. A. (2019). Evolutionary correlated gravitational search algorithm (ECGS) with genetic optimized hopfield neural network (GHNN)–a hybrid expert system for diagnosis of diabetes. *Measurement*, *145*, 551-558.

102. Kannadasan, K., Edla, D. R., & Kuppili, V. (2019). Type 2 diabetes data classification using stacked autoencoders in deep neural networks. *Clinical Epidemiology and Global Health*, *7*(4), 530-535.

103. El-Sappagh, S., Elmogy, M., Ali, F., & Kwak, K. S. (2019). A case-base fuzzification process: diabetes diagnosis case study. *Soft Computing*, *23*(14), 5815-5834.

104. Cho, S. B., Kim, S. C., & Chung, M. G. (2019). Identification of novel population clusters with different susceptibilities to type 2 diabetes and their impact on the prediction of diabetes. *Scientific reports*, *9*(1), 1-9.

105. Zhang, B., Hsu, M., & Dayal, U. (1999). K-harmonic means-a data clustering algorithm. Hewlett-Packard Labs Technical Report HPL-1999-124, 55.

106. Deng, D., & Kasabov, N. (2003). On-line pattern analysis by evolving self-organizing maps. *Neurocomputing*, *51*, 87-103.

107. Polat, K., Güneş, S., & Arslan, A. (2008). A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine. *Expert systems with applications*, *34*(1), 482-487.

108. Temurtas, H., Yumusak, N., & Temurtas, F. (2009). A comparative study on diabetes disease diagnosis using neural networks. *Expert Systems with applications*, *36*(4), 8610-8615.

109. Kayaer, K., & Yıldırım, T. (2003, June). Medical diagnosis on Pima Indian diabetes using general regression neural networks. In *Proceedings of the international conference on artificial neural networks and neural information processing (ICANN/ICONIP)* (Vol. 181, p. 184).

110. Carpenter, G. A., & Markuzon, N. (1998). ARTMAP-IC and medical diagnosis: Instance counting and inconsistent cases. *Neural Networks*, *11*(2), 323-336.

111. Dogantekin, E., Dogantekin, A., Avci, D., & Avci, L. (2010). An intelligent diagnosis system for diabetes on linear discriminant analysis and adaptive network based fuzzy inference system: LDA-ANFIS. *Digital Signal Processing*, *20*(4), 1248-1255.

112. Bozkurt, M. R., Yurtay, N., Yilmaz, Z., & Sertkaya, C. (2014). Comparison of different methods for determining diabetes. *Turkish Journal of Electrical Engineering & Computer Sciences*, *22*(4), 1044-1055.

113. Yilmaz, N., Inan, O., & Uzer, M. S. (2014). A new data preparation method based on clustering algorithms for diagnosis systems of heart and diabetes diseases. *J Med Syst*, *38*, 48-59.

114. Ramezani, R., Maadi, M., & Khatami, S. M. (2018). A novel hybrid intelligent system with missing value imputation for diabetes diagnosis. *Alexandria engineering journal*, *57*(3), 1883-1891.

115. Orkcu, H. H., & Bal, H. (2011). Comparing performances of backpropagation and genetic algorithms in the data classification. *Expert systems with applications*, *38*(4), 3703-3709.

116. Luukka, P. (2007). Similarity classifier using similarity measure derived from Yu's norms in classification of medical data sets. *Computers in Biology and Medicine*, *37*(8), 1133-1140.

117. Isa, N. A. M., & Mamat, W. M. F. W. (2011). Clustered-hybrid multilayer perceptron network for pattern recognition application. *Applied Soft Computing*, *11*(1), 1457-1466.

118. Aslam, M. W., Zhu, Z., & Nandi, A. K. (2013). Feature generation using genetic programming with comparative partner selection for diabetes classification. *Expert Systems with Applications*, *40*(13), 5402-5412.

119. Seera, M., & Lim, C. P. (2014). A hybrid intelligent system for medical data classification. *Expert Systems with Applications*, *41*(5), 2239-2249.

120. Van Belle, V., & Lisboa, P. (2014). White box radial basis function classifiers with component selection for clinical prediction models. *Artificial intelligence in medicine*, *60*(1), 53-64.

121. Wang, K. J., Adrian, A. M., Chen, K. H., & Wang, K. M. (2015). An improved electromagnetism-like mechanism algorithm and its application to the prediction of diabetes mellitus. *Journal of biomedical informatics*, *54*, 220-229.

122. Seera, M., Lim, C. P., Tan, S. C., & Loo, C. K. (2015). A hybrid FAM–CART model and its application to medical data classification. *Neural Computing and Applications*, *26*(8), 1799-1811.

123. Zhu, J., Xie, Q., & Zheng, K. (2015). An improved early detection method of type-2 diabetes mellitus using multiple classifier system. *Information Sciences*, *292*, 1-14.

124. Ding, S., Zhao, H., Zhang, Y., Xu, X., & Nie, R. (2015). Extreme learning machine: algorithm, theory and applications. *Artificial Intelligence Review*, *44*(1), 103-115.

125. Mohapatra, P., Chakravarty, S., & Dash, P. K. (2015). An improved cuckoo search based extreme learning machine for medical data classification. *Swarm and Evolutionary Computation*, *24*, 25-49.

126. Feng, T. C., Li, T. H. S., & Kuo, P. H. (2015). Variable coded hierarchical fuzzy classification model using DNA coding and evolutionary programming. *Applied Mathematical Modelling*, *39*(23-24), 7401-7419.

127. Luukka, P. (2011). Feature selection using fuzzy entropy measures with similarity classifier. *Expert Systems with Applications*, *38*(4), 4600-4607.

128. Polat, K., & Güneş, S. (2007). An improved approach to medical data sets classification: artificial immune recognition system with fuzzy resource allocation mechanism. *Expert Systems*, *24*(4), 252-270.

129. Polat, K., & Güneş, S. (2007). An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease. *Digital Signal Processing*, *17*(4), 702-710.

130. Ghazavi, S. N., & Liao, T. W. (2008). Medical data mining by fuzzy modeling with selected features. *Artificial Intelligence in Medicine*, *43*(3), 195-206.

131. D. Karaboga, An idea based on honey bee swarm for numerical optimization, Erciyes University, Kayseri, Turkey, Technical Report-TR06, 2005.

132. Dedeturk, B. K., & Akay, B. (2020). Spam filtering using a logistic regression model trained by an artificial bee colony algorithm. Applied Soft Computing, 106229.

133. Cao, Y., Lu, Y., Pan, X., & Sun, N. (2019). An improved global best guided artificial bee colony algorithm for continuous optimization problems. Cluster computing, 22(2), 3011-3019.

134. Li, J. Q., Song, M. X., Wang, L., Duan, P. Y., Han, Y. Y., Sang, H. Y., & Pan, Q. K. (2019). Hybrid Artificial Bee Colony Algorithm for a Parallel Batching Distributed Flow-Shop Problem With Deteriorating Jobs. IEEE transactions on cybernetics.

135. Sahoo, G. (2017). A two-step artificial bee colony algorithm for clustering. Neural Computing and Applications, 28(3), 537-551.

136. Venugopalan, J., Chanani, N., Maher, K., & Wang, M. D. (2019). Novel Data Imputation for Multiple Types of Missing Data in Intensive Care Units. IEEE journal of biomedical and health informatics, 23(3), 1243-1250.

137. Laña, I., Olabarrieta, I. I., Vélez, M., & Del Ser, J. (2018). On the imputation of missing data for road traffic forecasting: New insights and novel techniques. Transportation research part C: emerging technologies, 90, 18-33.

138. Jordanov, I., Petrov, N., & Petrozziello, A. (2018). Classifiers accuracy improvement based on missing data imputation. Journal of Artificial Intelligence and Soft Computing Research, 8(1), 31-48.

139. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18).

140. Gupta, S., Kumar, R., Lu, K., Moseley, B., & Vassilvitskii, S. (2017). Local search methods for k-means with outliers. Proceedings of the VLDB Endowment, 10(7), 757-768.

141. Aggarwal, C. C. (2012). A segment-based framework for modeling and mining data streams. Knowledge and information systems, 30(1), 1-29.

142. Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000, May). LOF: identifying density-based local outliers. In Proceedings of the 2000 ACM SIGMOD international conference on Management of data (pp. 93-104).

143. Li, X., Yang, Y., Pan, H., Cheng, J., & Cheng, J. (2019). A novel deep stacking least squares support vector machine for rolling bearing fault diagnosis. Computers in Industry, 110, 36-47.

PUBLICATIONS FROM THESIS

Journals

Accepted

1. Srivastava, Anand Kumar, Yugal Kumar, and Pradeep Kumar Singh. "Artificial Bee Colony and Deep Neural Network-Based Diagnostic Model for Improving the Prediction Accuracy of Diabetes." International Journal of E-Health and Medical Communications (IJEHMC), IGI Global, Vol.12, Issue 2, Article 2, 2020. (Scopus)
2. Srivastava, Anand Kumar, Yugal Kumar, and Pradeep Kumar Singh. "Computer aided diagnostic system based on SVM and K harmonic mean-based attribute weighting method." Obesity Medicine, Elsevier, Vol. 19, (2020): 100270. (Scopus)
3. Srivastava, Anand Kumar, Yugal Kumar, and Pradeep Kumar Singh. "A Rule-Based Monitoring System for Accurate Prediction of Diabetes." International Journal of E-Health and Medical Communications (IJEHMC), IGI Global, Vol. 11, no. 3 (2020): 32-53 (Scopus)
4. Srivastava, Anand Kumar, Yugal Kumar, and Pradeep Kumar Singh. "Hybrid Diabetes Disease Prediction Framework Based on Data Imputation and Outlier Detection Techniques", Expert System with Applications, Elsevier (Communicated, Submission Id: EXSY0001297). (SCI, Scopus)

Conference

1. Srivastava, Anand Kumar, Pradeep Kumar Singh, and Yugal Kumar. "A Taxonomy on Machine Learning Based Techniques to Identify the Heart Disease." In International Conference on Next Generation Computing Technologies, pp. 13-25. Springer, Singapore, 2018.
2. Srivastava, Anand Kumar, Pradeep Kumar Singh, and Yugal Kumar. "A Review on Recent Machine Learning Techniques for Diabetes Disease prediction', In International Conference International Conference on Inventive Engineering and Computing Techniques (ICIECT-2020) held on 17-18 June 2020 at Krishna Engineering College, Ghaziabad, UP, India.