ISSN: 2708-7123 | Volume-**01**, Issue Number-**03** | September-2020
LC INTERNATIONAL JOURNAL OF STEM
Web: www.logicalcreations.org/stem | www.lcjstem.com | DOI: https://doi.org/10.47150

A NOVEL FRAMEWORK FOR CLASSIFICATION OF DIABETES DISEASES PATIENTS USING RANDOM FOREST AS FEATURE SELECTION

Zeenat Bashir[1], Dr. Hamid Ghous[2]
[1,2]Dept. Computer science, Institute of Southern Punjab (ISP), Multan, Pakistan
[1]zeenatbashir16@gmail.com, [2]hamidghous@isp.edu.pk

ABSTRACT-Technological advancements are increasing day by day in biomedical field. A huge amount of data has been collected from different resources and used in the biomedical field. Data mining is a process of extracting the hidden patterns from large datasets to gain useful information for users. However, data mining trend in healthcare applications is increasing, and it is playing an important role in the medical field for diagnosing and predicting diseases at early stages. Data mining techniques such as classification, clustering, association, regression, and summarization have been previously used for diagnosing and predicting diseases. Diabetes is a common and chronic disease which causes an increase in blood sugar. Many complexities occur if diabetes remains unidentified and untreated. The present study aims are to implement Random Forest (RF) as a feature selection method and some classification method such as Support Vector Machine (SVM), Decision Tree (DT) and Neural Network on two diabetes dataset for the early diagnosis of diabetes. The proposed results show that the Support Vector Machine provides higher accuracy for the prediction of diabetes disease. It will be very effective and efficient for everyone.

Keywords- Diabetes, decision tree (DT), neural network (NN), random forest (RF), support vector machine (SVM).

I. INTRODUCTION

Problem Statement:

The data about diabetes patients is increasing day by day and factors causing diabetes disease increasing continuously. So, it is very difficult to diagnose diabetes disease in less time and effective manners. That's why; first this study used the random forest as a feature selections method to reduce the features of diabetes dataset after that, this research used classification methods to predict the diabetes disease. In this way, this research will reduce the time and cost as well as improve the prediction methods.

BACKGROUND

Diabetes Disease:

Diabetes is a very common disease these days in all age groups of people. According to 2016 report of World Health Organization (WHO), 422 million people have diabetes disease in 2014 and also expected that the ratio of diabetes patients rise over 380 million in 2025, Shuja Mirza & Dr. Sonumittal (2018), AiswaryaIyer et al. (2015).

Diabetes is the cause of heart disease, kidney disease, nerve damage, and blindness. So, it is a critical issue for mining diabetes efficiently and effectively. It affects the ability of the body in producing insulin. Diabetes is a disease in which a human body's glucose level increases. Glucose is much important for health, and it is a source of energy for cells, Ramin Ghorbania & Rouzbeh Ghousi **(**2019) Diabetes is a chronic disease. Intensify of thirst, hunger, and continual urination are symptoms caused due to high blood sugar. If diabetes disease remains unidentified and untreated it creates many complications. Diabetes affects many factors such as height, weight, heredity, insulin but the major factor considered is sugar among all factors. Early diagnosis of diabetes disease is the only remedy to stay away from complications, Deepti Sisodia & Dilip Singh Sisodia (2018).

ISSN: 2708-7123 | Volume-01, Issue Number-03 | September-2020
LC INTERNATIONAL JOURNAL OF STEM
Web: www.logicalcreations.org/stem | www.lcjstem.com | DOI: https://doi.org/10.47150

Diabetes is one of the dangerous diseases. It occurs when the desired amount of insulin is not produced which is required for the human body or when the human body cannot properly manage the produced insulin. The affected person's body cannot produce enough insulin, and that person will be unable to consume its insulin. Diabetes increases the sugar level in human blood. The cause of increasing the sugar level in blood is called diabetes as "sugar". Various symptoms are found in the affected person by the diabetes disease. Frequent urination, feeling pain in muscles, increased hunger, and thirsts are the main symptoms of diabetes. It needs early detection of disease so that risk level is decreased, Himansu Das et al., (2018). Diabetes Mellitus is the most growing disease that produced high blood sugar in the human body. Diabetes affects the human body to utilize the sugar which is present in food. Type 1 diabetes, Type 2 diabetes, and Gestational diabetes are three different types of diabetes diseases. All types of diabetes need to be predicted at its early stage as it is a lifelong disease, and there is no cure for it. We can control it at an early stage only, Amina Azrar et al., (2018).

Types of Diabetes:

Three types of diabetes are defined below:

1. **Type 1 diabetes:**

Pancreas does not produce accurate amounts of insulin in this type of diabetes. People which have this type of disease depend on external injected insulin to maintain the glucose level in the body. Genetic factors are the causes of this type of disease.

2. **Type 2 diabetes:**

Insulin resistance occurs in this type of disease and the body cannot use insulin properly. Overweight people caused this type of disease. It mostly affects the heart. Heart diseases and heatstroke are common causes of this type of diabetes. It can only be control with proper treatment.

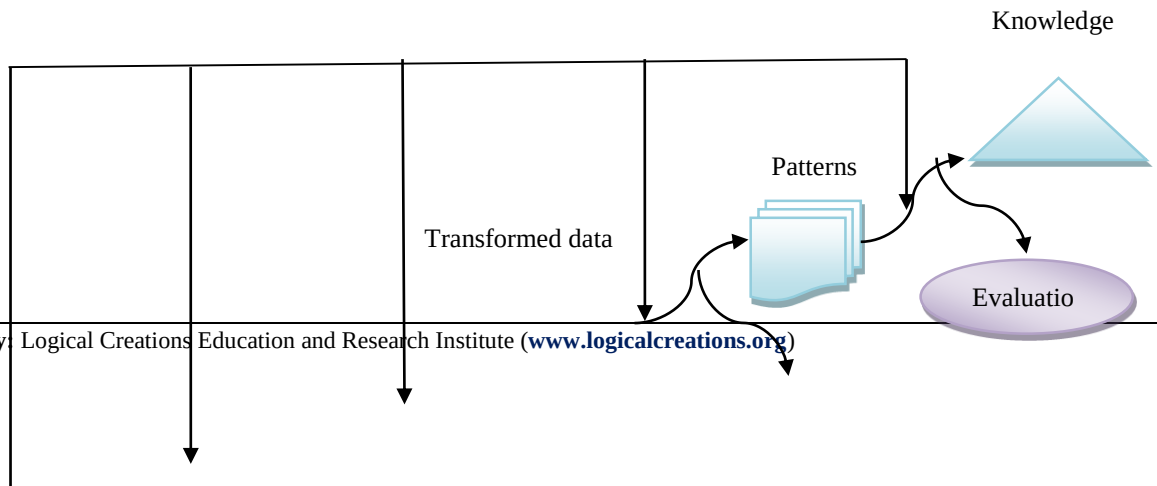
3. **Gestational diabetes:**

Married women are affected with this type of diabetes. During pregnancy insulin blocking hormones are produced which affects pregnant women. High blood sugar is the cause of this type of diabetes, Amina Azrar et al., (2018).

It is reported that the effect of diabetes has a more fatal and worsening impact on women than on men because of their lower survival rate and poorer quality of life. WHO report stated that, almost one- third of the women who suffer from diabetes have no idea about it. The effects of diabetes are unique in the case of mothers because diabetes disease is transmitted to their unborn child. Strokes, miscarriages, blindness, kidney failure, and amputations are just some complications that arise from this disease, Aiswarya Iyer et al., (2015). A person is considered as suffering from diabetes when his blood sugar level increases. A diabetic patient's body is not able to produce or use insulin well. Type 1, Type 2, and Gestational are three types of diabetes disease. All types of diabetes disease are dangerous and need treatment. One can avoid the complications related to them, if these are detected in the early stages, Aiswarya Iyer et al., (2015).

Knowledge discovery in databases (KDD)

It is the process of attaining useful knowledge or information from the huge collection of data. All steps of knowledge extraction are significant to obtain meaningful information. Fig.1: define the knowledge discovery steps, Ramin Ghorbani & Rouzbeh Ghousi, (2019):



ISSN: 2708-7123 | Volume-01, Issue Number-03 | September-2020
LC INTERNATIONAL JOURNAL OF STEM
Web: www.logicalcreations.org/stem | www.lcjstem.com | DOI: https://doi.org/10.47150

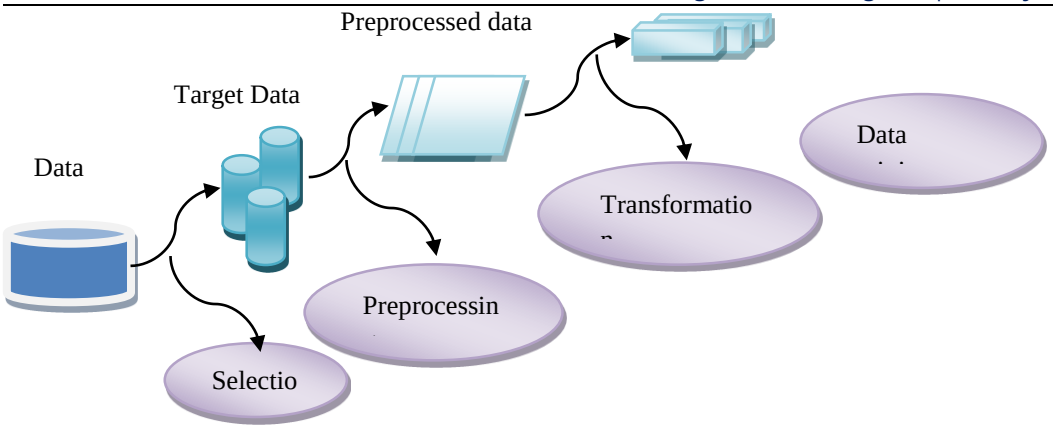


Fig. 1: Steps of knowledge discovery in diabetes

Data Mining

It is a step of knowledge discovery in database which is used to collect useful information. Data mining (DM) is a process of analyzing and selecting hidden pattern for obtaining useful information. Data mining have different application in which one of them is medical diagnosis. Today many diseases such as heart disease, breast cancer and diabetes disease are the most dangerous ones. Many data mining techniques have been applying for diagnosing and predicting diseases such as classification, clustering and association rules. For data analyzing classification techniques is known best. Bayesian network, Artificial neural network, decision tree, support vector machine, K-Nearest neighbor, Associative classification, Rule-based classification, Genetic algorithm, Rough set approach and Fuzzy set classification are classification methods. Clustering methods are used to find the similar data which are related to one another in the form of clusters. Clustering techniques specified classes and objects in each category, while classification techniques specified objects in predefined category. Association rule find the new relations among variables in database. It also used to find the patterns between collections of items, Ramin Ghorbani & Rouzbeh Ghousi (2019).

Data mining is an emerging field with the vast variety of techniques from different field. Data mining is a combination of statistics, machine learning, pattern recognition and artificial intelligence system. It is used to analyze huge amount of data to discover the hidden patterns in the data. It also used in medical sciences for decision-making and to obtain hidden knowledge from a huge amount of diabetes data which provides the quality treatment for diabetes suffering patients, Himansu Das et. al, (2018).

Primary step of data mining include selection of data, preprocessing data, transformation of data, mining data, and last evaluation of pattern and recognition of pattern. It is a process of obtaining meaningful information from any dataset. Techniques used for data mining are association rule, classification, clustering. Various rules implemented using data mining techniques. It is used for predicting diseases. Selecting disease a lot of records required such as affected patient's history, hospitals, clinical devices and electronic facts. These records are used for selecting useful information in which we are able to take options and generate different rules. Different diseases are diagnosed using data mining techniques, for example, AIDS, diabetes, heart disease and breast cancer, Amina Azrar et al., (2018).

Large amount of information is gathered in the form of patient's record from hospitals. Prediction purpose is done through data mining. This method helpful in decision-making through algorithms we extract useful information from a huge amount

ISSN: 2708-7123 | Volume-01, Issue Number-03 | September-2020
LC INTERNATIONAL JOURNAL OF STEM
Web: www.logicalcreations.org/stem | www.lcjstem.com | DOI: https://doi.org/10.47150

of data, which is collected from medical centers. Data mining techniques applied in detection of diabetes at its early stage and treatment. It is helpful in avoiding complications, Aiswarya Iyer et al., (2015).

CLASSIFICATION

In this research we used classification methods for predicting diabetes disease in patients. To classify the data this study implemented three classification methods on diabetes disease patients which are as follows:

I. Support Vector Machine (SVM):

SVM is a supervised machine learning method which is used for classification. The objective of the support vector machine is to find out the best hyper-plane between two classes. The best hyper-plane should not be closer to the other class data points. From each category those hyper-plane are selected which are far from data points. Those data points which are closer to the margin of the classifier are called support vector, Deepti Sisodia & Dilip Singh Sisodia (2018).

II. Decision Tree:

Supervised machine learning algorithms used decision tree for solving classification problems. For prediction and classification, decision tree used nodes and internodes. The parent node consists of two or more than two branches; on the other hand, the leaf node defines the classification. From all attributes, decision trees select every node for obtaining the highest information, Deepti Sisodia & Dilip Singh Sisodia, (2018).

III. Neural Network:

Interconnected neurons called a neural network. The input layer, hidden layer, and output layer are the three types of neural networks. It works as a brain and consists of various neurons. So, it is called a neural network, Imola K. Fodor, (2002).

1.1 Feature Selection

In this research we used Random Forest (RF) as a features selection method.

The feature selection technique is used to improve the efficiency of data mining algorithms. It is a process to remove irrelevant and redundant information. It reduced the attribute which is not useful in the dataset. Various attributes available in the database, but only useful attributes are used. Noisy, irrelevant, and redundant feature in data is a big problem in the world. Feature selection clearly removed irrelevant data for diagnosing diabetes disease, Yue Huang et al., (2004).

1.2 Identify Research questions:

• Can we use the random forest as a feature selection method for diabetes disease data?

• Can a framework using data mining techniques help biologists in early diagnosis of diabetes?

• Is this model time-effective and cost-effective also?

1.1 Significance/Objective/Scope:

Diabetes is the most common disease in every age group. It contains different types such as type I diabetes, type II diabetes, etc. Nearly half of all diabetics have heredity factors. This disease is transformed from parents to children through gene transferred. In 2004, an expected 3.4 million people passed on from the result of fasting high blood sugar. Some studied tells that nearly 98,000 people died each year, Ahmed Hamza Osman & Hani Moetque Aljahd ali, (2017).

According to the 1999 World Health Organization (WHO), the total number of 1487 people at the age of 20 years and older are included. A total of 735 people were confirmed to have diabetes and pre-diabetes. Remaining 752 people were not diabetes or pre-diabetes patients and were confirmed as such by physical check-in past two years, Xue-Hui Meng, (2013).

The purpose of this study is to help biologists to diagnose diabetes at early stage. In this study, we will try to use the random forest as a feature selections method and then implement different classification methods to find out the best method among them. In this way, our work will be helpful for the community to predict diabetes disease at an early stage.

1.2 Summary

ISSN: 2708-7123 | Volume-**01**, Issue Number-**03** | September-2020
LC INTERNATIONAL JOURNAL OF STEM
Web: www.logicalcreations.org/stem | www.lcjstem.com | DOI: https://doi.org/10.47150

This chapter provides a brief introduction and background history of work. It also provides the background about feature selection methods and classification techniques. It defines the problem statement, objectives/Scope/Significance of this work. This chapter identifies the research questions of this study. In the next chapter, the literature will be reviewed which is related to classification methods and feature selection methods explained in this chapter.

II. LITERATURE REVIEW

In the past, many researchers had worked on early diagnosis of diabetes disease. Diabetes disease is one of the chronic diseases and becoming a cause of death in people. So many factors involved which cause diabetes disease and in this way, a huge amount of data increasing about diabetes disease. That's why researchers have been working to handle this huge amount of data. Different studies have been carried out by using data mining techniques like Decision tree, Support vector machine, Naïve Bayes, Neural network and random forest, etc. All methods show the performances of these models for the diagnosis of diabetes disease.

Deepti Sisodia & Dilip Singh Sisodia (2018) designed a model for diagnosing diabetes in patients with maximum accuracy. They used three machine learning-based classification techniques such as support vector machine, Decision tree, and Naive Bayes algorithm to predict early-stage diabetes. These experiments are performed on the Pima Indians Diabetes Database (PIDD) which is collected from the UCI machine learning repository. The results of all three algorithms are calculated on different measures like Precision, Accuracy, F-Measures, and Recall. Results show that the Naïve Bayes algorithm provides higher accuracy of 76.30% than the other two algorithms. SVM provides a minimum accuracy of 65.10%. These accuracy results are measured with Receiver Operating Characteristic curves properly and systematically.

Ahmed Hamza Osman & Hani Moetque Aljahd ali (2017) proposed an integrated approach for diagnosing diabetes disease using the Support vector machine and K-Means clustering algorithms. They used the UCI Pima Indian diabetes standard dataset. In the dataset, they select a useful attribute for improving classification accuracy. Based on SVM and K-Means, a T-test statistical method achieved improved results. Integration between K-Means and SVM can enhance the diagnosis results in diabetes disease.

Amina Azrar et al., (2018) presented data mining algorithms for diagnosing early-stage diabetes disease. They provide a comparison and show the best algorithm for prediction in healthcare fields. Decision Tree, Naïve Bayes, and KNN algorithms are used for diagnosing the early stage of diseases. Each of these algorithms can give high accuracy and efficiency depending upon the type of data and attributes. After the Experimentation of these three algorithms, it can be said that for Pima Indian diabetes (PID) dataset Decision Tree shows the best accuracy which is higher than KNN and Naïve Bayes algorithms. The tool used for testing is Rapid Miner.

Rahul Joshi1 &Minyechil Alehegn (2017) used four classification algorithms such as KNN, Naïve Bayes, random forest, and J48 to classify a diabetes patient. This study aims to classify the diabetes disease and compare the all algorithm's accuracy. Patient analyses with positive and negative values on the basis of some measurements. The highest performance algorithm provides the best method to diagnose the diabetes disease at its early stage. They also proved that the single algorithm gives less accuracy than a hybrid method. The result shows that the decision tree provided higher accuracy from the other three algorithms for predicting diabetes analysis.

Emrana Kabir Hashi et al., (2017) diagnosing diabetes disease using data mining techniques. They proposed a system for predicting disease which is helpful for physicians, doctors, medical students, and patients for deciding diagnosing disease. In this system, they used 70:30 percentages for train and test data. In the training phase, both algorithms provide 100% accuracy but in the test phase, KNN gives 90.43% and C4.5 gives 76.96% accuracy. Decision tree and KNN algorithms are used for calculation and comparing the accuracy of

ISSN: 2708-7123 | Volume-**01**, Issue Number-**03** | September-2020
LC INTERNATIONAL JOURNAL OF STEM
Web: www.logicalcreations.org/stem | www.lcjstem.com | DOI: https://doi.org/10.47150

experimental results. The result shows that the decision tree gives better accuracy of 90.43% for diagnosing diabetes disease.

N. Sneha1 &Tarun Gangil (2019) used predictive analysis for early detection of diabetes mellitus. The objective of this work is to analyze the diabetes dataset using some classification methods and reduce the complexity of diabetes prediction. It improves the prognosis of diabetes people. Performance measures are based on sensitivity, specificity, recall, and precision. Five classification methods such as Decision tree, Random forest, KNN, Support vector machine, Naïve Bayes algorithms used for the prediction of diabetes disease. The result proved that the Support vector machine provides higher accuracy of 77.73% for diagnosing early-stage diabetes mellitus and KNN provides a minimum accuracy of 63.04%.

Ratna Nitin Patil& Dr. Sharvari Chandra shekhar Tamane (2018) presented a framework for the detection of diabetes. They used feature selection techniques such as k-nearest neighbor and naïve Bayes approach to developing a proposed model which diagnoses the patient is diabetic or none. The main objective of feature selection is to reduce the features which are used in classification for obtaining higher accuracy. Genetic Algorithm (GA) for feature selection is used to remove the redundant or irrelevant features for mining the best accuracy. PIMA Indian diabetes dataset be used for analysis. The proposed model is compared with traditional models. The result shows that the new model gives higher accuracy than earlier models.

Sofia Benbelkacem& Baghdad atmani (2019) focused on random forest classification method for diagnosing chronic diabetes disease. Different numbers of trees were developed based on random forest. After that, it compared with other machine learning algorithms. Five supervised learning algorithms, C4.5, REPTree, SimpleCart, BFTree, and SVM compared with random forest for obtaining accuracy. The Pima Indian diabetes dataset used which is selected from the UCI repository for this experiment. Two stages are defines in experiments, in the first stage they select a random number of trees from the random forest and test the accuracy with the different number of trees. In the second stage, compare the algorithms with other machine algorithms. Accuracy is evaluated based on sensitivity and specificity. The result shows that the random forest proved more efficient than other machine algorithms.

Seyed Ataaldin Mahmoudinejad et al. (2019) proposed a new ensemble model based on data mining methods for early diagnosis of diabetes disease. They used a weighted k-nearest neighbor, decision tree, and logistic regression classification methods for preprocessing. For diagnosis diabetes mellitus they evaluated different various diabetes risk factors. In this research Pima Indian diabetes dataset was used which is collected from the UCI repository. The dataset consists of eight attributes and 768 instances. For controlling data scattering and improving classification results author's used the data normalization method. They also implemented 10-cross validation for estimation of error rate, and classification performance based on training and testing datasets. In this study, the proposed hybrid model compares with single classifiers, and the result shows that the new ensemble model provides higher accuracy of 80.60% than other all classifiers. It also described that the hybrid model always provides higher accuracy than other all single classifiers.

N. Yuvaraj & K.R. Sri Preetha (2019) proposed a new model for predicting diabetes disease. They used three different machine learning (ML) algorithms such as random forest, decision tree, and naïve Bayes. Pima Indian diabetes dataset is used which is sourced from the National Institute of Diabetes and Digestive Diseases after preprocessing of data. The dataset consists of thirteen attributes and 75,664 instances. For the extraction of useful information, they applied feature selection methods for reducing noise and irrelevant features. Information Gain (IG) method is used as a feature selection method. From thirteen attributes they selected only eight attributes from the dataset. The performance was evaluated based on 70% training

ISSN: 2708-7123 | Volume-**01**, Issue Number-**03** | September-2020
LC INTERNATIONAL JOURNAL OF STEM
Web: www.logicalcreations.org/stem | www.lcjstem.com | DOI: https://doi.org/10.47150

and 30% testing of the dataset. The result showed that the Random forest provides higher accuracy of 94%, Naïve Bayes of 91%, and a decision tree of 88%.

Francesco Mercaldo et al. (2017) presented a new model for the classification of diabetes. They used six different classifiers including J48, Random forest, JRip, HoeffdingTree, Multilayer perceptron, and BayesNet. Pima Indian diabetes dataset was used in this research which consists of eight attributes and 768 instances. Female patients are tested in this dataset which is 21 years old. The researchers implemented two algorithms, GreedStepwise and Best First for determining those attributes which are the helpful increasing performance of classification. Only four attributes were selected from the dataset for testing diabetes patients. The four attribute names are body mass index, diabetes pedigree function, plasma glucose concentration, and age. Accuracy measures are based on 10-cross validation, precision, recall, and F-measures. Using the Hoeffding Tree algorithm result evaluated that the precision value is 0.757, recall value 0.762, and F-measures value is 0.759 which is the higher accuracy of performance than others all.

Ambika Choudhury & Deepak Gupta (2019) used six machines learning techniques for early diagnosis of diabetes such as Support vector machine (SVM), Logistic regression, Naïve Bayes (NB), Random forest (RF), Decision tree (DT) and Artificial neural network (ANN). Pima Indian diabetes dataset was utilized for this experiment which consists of 768 instances and nine attributes. Comparison of research evaluated based on specificity, precision, recall, accuracy, false-positive rate (FP rate), and negative predictive value (NPV), G-means, and F-measures. All classification methods, performance analyzed in terms of accuracy rate. The result showed that logistic regression provides higher accuracy for the prediction of diabetes disease. It provides an accuracy of 0.7761 which is higher than previous researches.

Ali Kalantari et al. (2018) implemented Computational Intelligence (CI) methods for evaluating single and hybrid methods incorrect prediction diseases based on accuracy, sensitivity, and specificity. Researchers used the Pima Indian diabetes dataset which is collected from the University of California at Irvine (UCI) repository for the experiment. They utilized different single and hybrid methods for the classification of diseases. Single methods are fuzzy logic, Genetic algorithms (GA), Particle swarm optimization (PSO), artificial neural network (ANN), Kernel method (KM) (support vector machine), and artificial immune system (AIS). Hybrid methods are Neuro-fuzzy (ANN, Fuzzy logic), Fuzzy support vector machine (FSVM), Fuzzy and genetic algorithm (FGA), Artificial immune system and generic algorithm (AIS-GA), Artificial immune system and neural network (AIS-NN), Particle swarm optimization and genetic algorithm (PSO-GA), SVM-GA, SVM-AIRS. After comparison, the result evaluated that the single methods provide the best accuracy of prediction in medical applications but hybrid model give the highest accuracy in term of accuracy, sensitivity, and specificity. Hybrid model Support vector machine with an artificial immune recognition system (SVM-AIRS) achieved 100% accuracy than other hybrid methods.

Minyechil Alehegn et al. (2018) used different classification methods to predict diabetes. Decision stump (DS), Naïve Net (NN), Support vector machine (SVM), and proposed ensemble method (PEM) implemented on Pima Indian diabetes dataset which is collected from UCI repository, and it consists of 768 instances and eight attributes. Preprocessing of data is required in this experiment for missing values and removes duplication of data. The proposed Ensemble Method (PEM) method means, combining the individual methods to make a hybrid model. After developing a hybrid model it increases the accuracy rate for the prediction of diabetes disease. 10-cross validation applied for evaluation of prediction performance. In the end comparison of the proposed method and individual method done and result obtained that the proposed model provides higher accuracy of 90.36% and decision stump provides a minimum accuracy of 83.72%.

ISSN: 2708-7123 | Volume-**01**, Issue Number-**03** | September-2020
LC INTERNATIONAL JOURNAL OF STEM
Web: www.logicalcreations.org/stem | www.lcjstem.com | DOI: https://doi.org/10.47150

N. Komal Kumar et al. (2019) aimed to develop a hybrid model using an optimized random forest classifier with a genetic algorithm for predicting diabetes disease. The diabetes dataset used in this experiment is collected from the University of California at Irvine which consists of fifty attributes and more than lack sample patients. Preprocessing required in this experiment for reducing irrelevant values and normalization of data. After preprocessing the samples remained in 2000 which is based on training and testing dataset, the performance of classification was obtained on the basis of accuracy, sensitivity, specificity, and kappa statistics. The result compared with existing hybrid classifier models and achieved a higher accuracy of the proposed hybrid model. In this study, the proposed hybrid model of the Genetic Algorithm with Optimized Random Forest classifier (GA-ORF) provides an accuracy of 0.923, sensitivity of 0.901, specificity of 0.924, and kappa statics of 0.879 which are higher from previous all researches for diabetes prediction.

2.1 Research Gap

This section describes the previous work of various researchers, and also describes the novelty of this project that how this model different from past studies.

Previous Work:

This chapter consists of a comparative analysis of the previous research. In previous work, various feature selection methods were used on different diabetes datasets to select some significant features from a large number of features. The researchers also processed diabetes datasets by removing irrelevant features and by eliminating redundant features. This process is called the preprocessing of data. After preprocessing of data, researchers applied different data mining techniques to predict the diabetes disease at its early stage and improve the accuracy of different prediction methods.

Novelty of this Work:

In contrast to previous researches, this study used two diabetes disease dataset; one is Pima Indian diabetes dataset 1 (UCI Repository) and the other is Pima Indian diabetes dataset 2 (Kaggle dataset). As we know, factors causing diabetes disease increasing from some decades. So, it was a very difficult task to predict diabetes disease. This was also very costly and time-consuming.

That's why; in this proposed research first we reduce the features of these datasets by implementing the Random Forest (RF). After preprocessing of datasets, this study applied some classification methods such as Support Vector Machine (SVM), Decision Tree (DT), and Neural Network (NN). This research leads to predict diabetes disease in a short time, and within a cost-effective manner. This project also improves the performance of the prediction model by increasing the classification accuracy. The proposed study helps the biologist to use feature selection methods for the largest datasets to early diagnosis of diabetes disease.

2.1 Summary

This chapter provides a detailed literature review about data mining, feature selection, and classification methods. The literature review also defines the research gap in current research and novelty of this work and that how the proposed model different from past studies. Brief introduction and background of the study, problem statement, and significance/objective/scope and identifies research questions explain in the previous Introduction chapter. In the next chapter, the proposed framework defines feature selection method and classification methods.

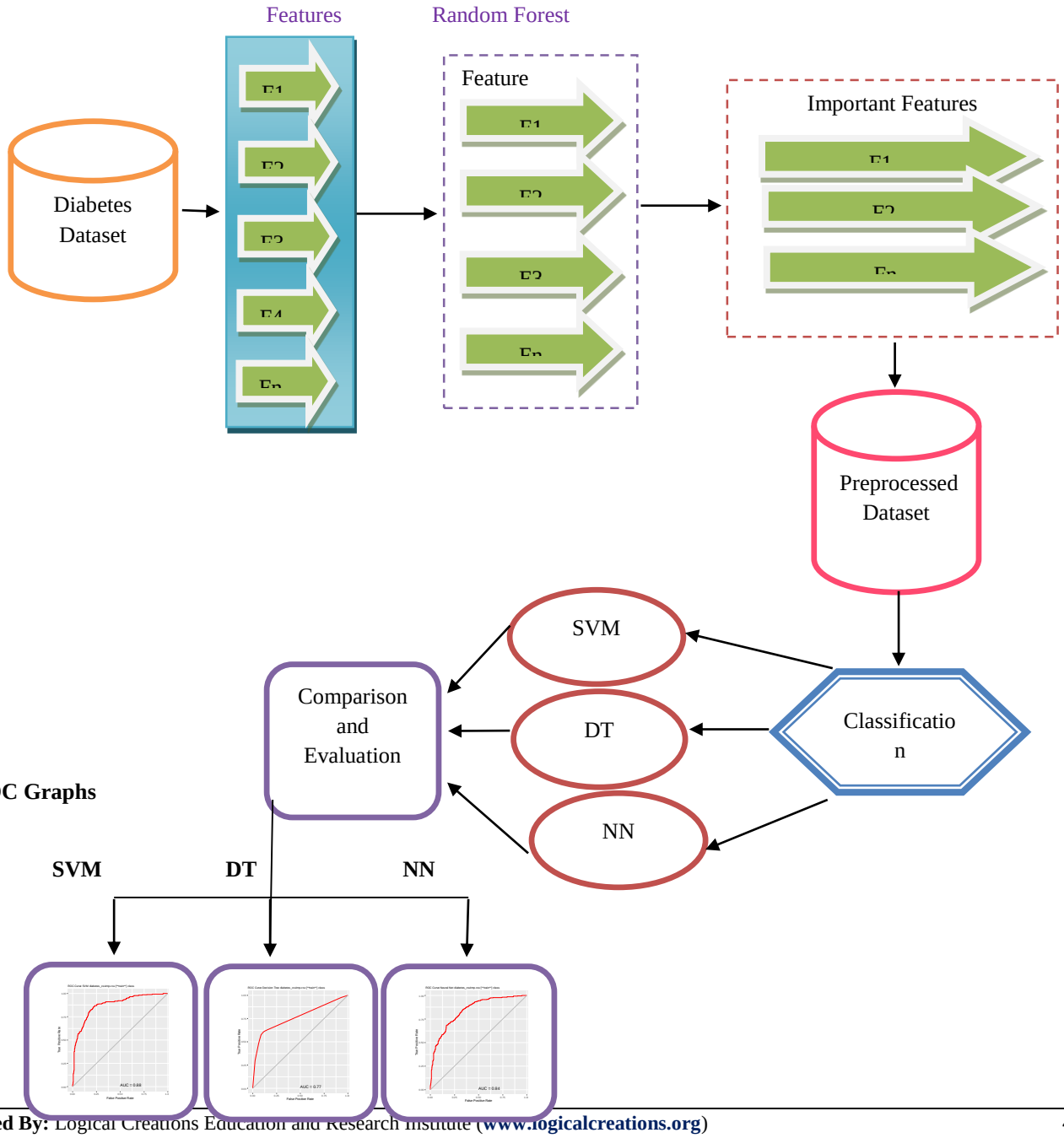
III. METHODOLOGY

This study has used two datasets from two different websites. The objective of these datasets is to predict whether a patient has diabetic or non-diabetic, on the bases of specific diagnostic measurements which are included in the dataset. This dataset consists of 768 instances which are female patients. Tested positive instances are 268 and the remaining 500 instances are tested negative. In target class variable Tested positive or tested negative shows whether the patient is diabetic or not. The first dataset also consists of numeric valued 8 attributes and the

ISSN: 2708-7123 | Volume-01, Issue Number-03 | September-2020
LC INTERNATIONAL JOURNAL OF STEM
Web: www.logicalcreations.org/stem | www.lcjstem.com | DOI: https://doi.org/10.47150

second dataset consists of 9 attributes in which 0 and 1 value used for testing diabetes. The value '0' tested for negative diabetes and value '1' for positive diabetes. In this study, Random Forest used as a feature selection method. After the preprocessing of data, this research applied some classification methods such as Support vector machine, Decision tree, and Neural Network, to predict the diabetes disease. All these experiments were implemented by using the R programming machine learning language. The complete frame-work of the proposed model is shown in Fig.2:

Model Diagram:



AU-ROC Graphs

SVM **DT** **NN**

ISSN: 2708-7123 | Volume-01, Issue Number-03 | September-2020
LC INTERNATIONAL JOURNAL OF STEM
Web: www.logicalcreations.org/stem | www.lcjstem.com | DOI: https://doi.org/10.47150

Fig.2: Complete framework of proposed model

3.1 Summary

In this chapter, first of all, explain both datasets its number of instances and attributes. After that, explain the flow of the proposed experiment on the basis of the diagram. The proposed model shows how features are selected with random forest and how classification methods work for obtaining higher accuracy. In the next chapter, the data analysis procedure and results show which is explained in this chapter in the form of the proposed model.

IV. DATA ANALYSIS PROCEDURE & RESULTS

In this proposed study R software is used for interpretation and analysis of data. It is an open-source programming language tool which is used for graphics and statistical computing. R tool provides factual methods and graphical representation which are non-linear and linear techniques, old-style factual tests, time arrangement investigation, order bunching and are exceptionally extensible. Extensibility and magnificent information representation are the two fundamental explanations behind the accomplishment of R.

- **Simulation of dataset 1(UCI Repository dataset) without training and testing**

First we have applied Random Forest (RF) as a feature selections method on the UCI Repository diabetes dataset. After selecting distinct features we have applied three classification methods as Support vector Machine (SVM), Decision Tree (DT), and Neural Network (NN). We have calculated Area under the Receiver Operating Curve (AU-ROC), Error Rate, and Time to built-in seconds of these three classification methods. Support Vector Machine (SVM) in these classifications methods shows the highest AU-ROC which is 0.87; Neural Network (NN) shows 0.84, and Decision tree (DT) shows 0.76 given in TABLE 1:

MODEL	ROC-ACCURACY %	ERROR RATE %	TIME IN SECONDS
SVM	0.8778	24	0.20
NN	0.8423	27.3	0.23
DT	0.7653	25.55	0.03

TABLE 1: SIMULATION OF DATASET 1 (UCI REPOSITORY DATASET) WITHOUT TRAINING AND TESTING.

- **Simulation of 20% testing on Dataset 1 (UCI Repository dataset)**

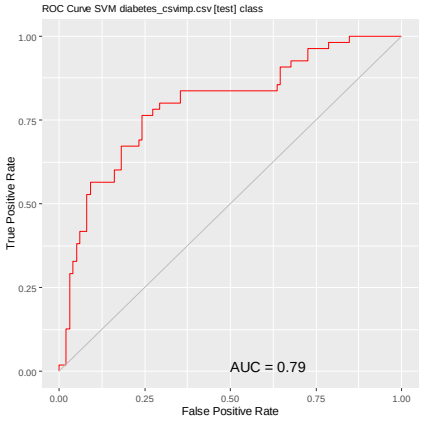
We have also applied these three classification methods support vector machine (SVM), Decision tree (DT), and Neural Network (NN) on UCI Repository dataset with 20% testing and 80% training after extracting important features. We have calculated Area under the Receiver Operating Curve (AU-ROC), Error rate, and time to built-in seconds of these three classification methods based on the testing dataset. Support Vector Machine in this classifications methods shows AU-ROC which is 0.79, Decision tree shows 0.774; and Neural network shows 0.776 given in TABLE 2:

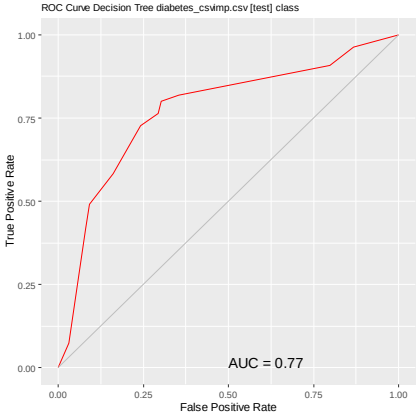
MODEL	ROC-ACCURACY %	ERROR RATE %	TIME IN SECONDS
SVM	0.79	26.85	1.73
DT	0.77	29	0.03
NN	0.78	26.55	0.16

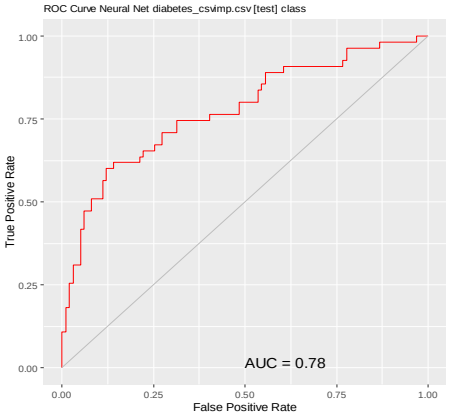
TABLE 2: SIMULATION OF DATASET 1 (UCI REPOSITORY DATASET) WITH 20% TESTING.

ISSN: 2708-7123 | Volume-01, Issue Number-03 | September-2020
LC INTERNATIONAL JOURNAL OF STEM
Web: www.logicalcreations.org/stem | www.lcjstem.com | DOI: https://doi.org/10.47150

ROC-Curves







Graph 1: Support Vector Machine (SVM) **Graph 2: Decision Tree** **Graph 3: Neural Network**

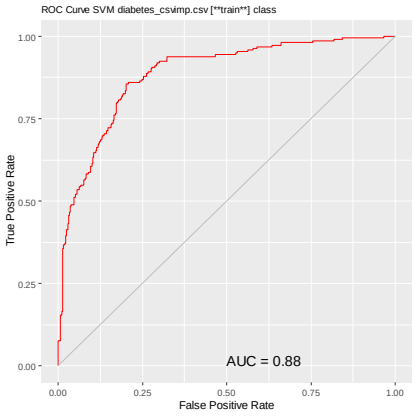
- **Simulation of 80% training Dataset 1 (UCI Repository dataset)**

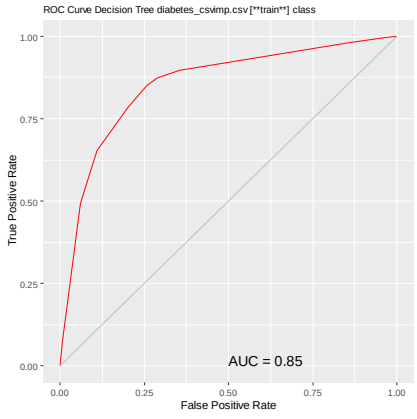
Then we have applied these three classification methods Support vector machine (SVM), Decision Tree (DT), and Neural Network (NN) on UCI Repository dataset with 20% testing and 80% training after implementing Random forest. We have calculated Area Under the Receiver Operating Curve (AU-ROC), Error rate, and time to built-in seconds of these three classification methods based on training. The support vector machine in these classifications methods shows the highest AU-ROC which is 0.88, Neural Network shows 0.86; and the Decision tree shows 0.85 given in TABLE 3:

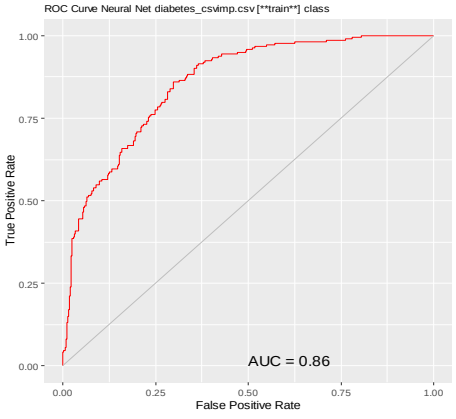
MODEL	ROC-ACCURACY %	ERROR RATE %	TIME IN SECONDS
SVM	0.88	24.35	0.16
DT	0.85	22.85	0.03
NN	0.86	27.15	0.16

TABLE 3: SIMULATION OF DATASET 1 (UCI REPOSITORY DATASET) WITH 80% TRAINING.

ROC-Curves







Graph 4: Support Vector Machine (SVM) **Graph 5: Decision Tree (DT)** **Graph 6: Neural Network (NN)**

- **Simulation of Dataset 2 (kaggle dataset) without training and testing**

We have also applied these three classification methods Support Vector Machine (SVM), Decision Tree (DT), and

LC-JSTEM
Your Research Partner
ISSN: 2708-7123 | Volume-**01**, Issue Number-**03** | September-2020
LC INTERNATIONAL JOURNAL OF STEM
Web: www.logicalcreations.org/stem | www.lcjstem.com | DOI: https://doi.org/10.47150

Neural Network (NN) on the Kaggle dataset without training and testing after implementing the Random forest. We have calculated Area under the Receiver Operating Curve (AU-ROC), Error Rate, and time to built-in seconds of these three classification methods. Support Vector Machine (SVM) in these classifications methods shows the highest accuracy which is 0.83, Decision Tree (DT) shows 0.74; and Neural Network shows 0.50 given in TABLE 4:

MODEL	ROC-ACCURACY %	ERROR RATE %	TIME IN SECONDS
SVM	0.831	27.7	0.23
DT	0.749	28.1	0.02
NN	0.500	50	0.03

TABLE 4: SIMULATION OF DATASET 2 (KAGGLE DATASET) WITHOUT TRAINING AND TESTING.

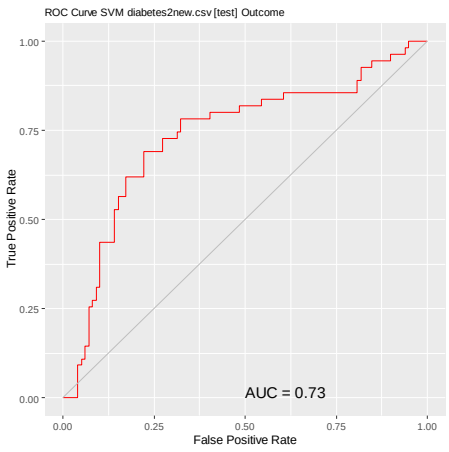
- **Simulation of 20% testing on Dataset 2 (Kaggle dataset)**

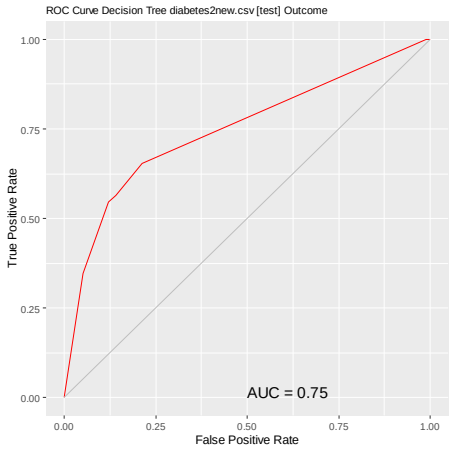
We have also applied these three classification methods support vector machine (SVM), Decision tree (DT), and Neural Network (NN) on the Kaggle dataset with 20% testing and 80% training after extracting important features. We have calculated Area under the Receiver Operating Curve (AU-ROC), Error rate, and time to build in seconds of these three classification methods based on the testing dataset. Support Vector Machine in these classifications methods shows AU-ROC which is 0.73, Decision tree shows 0.74; and Neural network shows 0.50 given in TABLE 5:

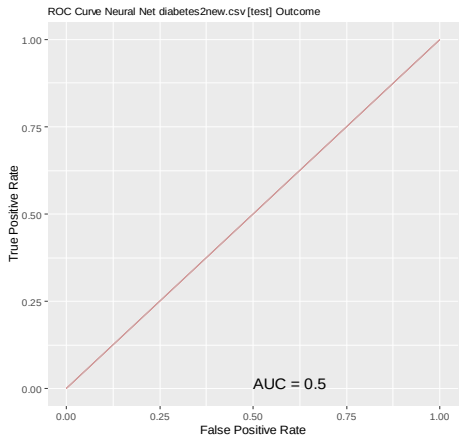
MODEL	ROC-ACCURACY %	ERROR RATE %	TIME IN SECONDS
SVM	0.73	32.5	0.31
DT	0.75	28.85	0.02
NN	0.50	50	0.03

TABLE 5: SIMULATION OF DATASET 2 (KAGGLE DATASET) WITH 20% TESTING.

ROC-Curves







Graph 7: Support Vector Machine (SVM) **Graph 8: Decision Tree (DT)** **Graph 9: Neural Network (NN)**

- **Simulation of 80% training on Dataset 2 (Kaggle dataset)**

We have applied these three classification methods Support vector machine (SVM), Decision Tree (DT), and Neural Network (NN) on UCI Repository dataset with 20% testing and 80% training after implementing Random forest. We have calculated Area under the Receiver Operating Curve (AU-ROC), Error rate, and time to built-in seconds of these three classification methods based on the training. The support vector machine in these classifications methods shows the highest AU-

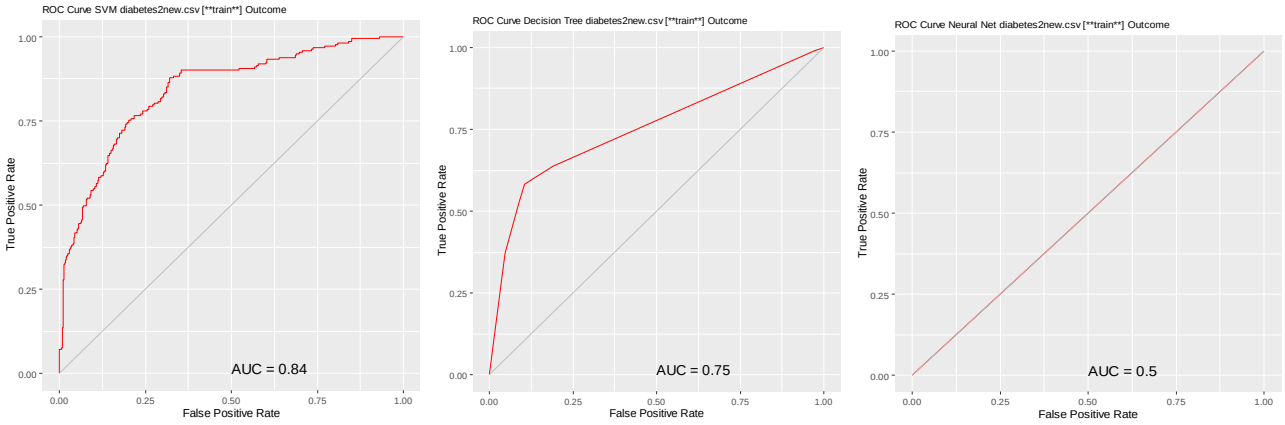
ISSN: 2708-7123 | Volume-01, Issue Number-03 | September-2020
LC INTERNATIONAL JOURNAL OF STEM
Web: www.logicalcreations.org/stem | www.lcjstem.com | DOI: https://doi.org/10.47150

ROC which is 0.84, Neural Network shows 0.50; and the Decision tree shows 0.75 given in TABLE 6:

MODEL	ROC-ACCURACY %	ERROR RATE %	TIME IN SECONDS
SVM	0.84	27.6	0.31
DT	0.75	26.15	0.02
NN	0.50	50	0.03

TABLE 6: SIMULATION OF DATASET 2 (KAGGLE DATASET) WITH 80% TRAINING.

ROC-Curves



Graph 10: Support Vector Machine (SVM) **Graph 11: Decision Tree (DT)** **Graph 12: Neural Network (NN)**

In this research work, implement the Random Forest (RF) as a feature selections method, with the combination of classification methods to diagnose the diabetes disease. This study concludes that, by reducing features, the performance of the prediction model is improved. This research calculated the highest AU-ROC by using a support vector machine (SVM) with 80% training on both datasets.

- **Comparison with other researcher's experiments.**

We compared our propose research results with previous researcher's experiments using the same type datasets, and we obtain, that our proposed model accuracy achieved higher level improvement.

TABLE 7: Define the comparisons of different accuracy methods between previous researchers. First-line results show the higher accuracy of the proposed research.

TABLE 7: COMPARISON BETWEEN PREVIOUS RESEARCHER'S EXPERIMENTS:

REFRENCES	YEAR	TECHNIQUES	ACCURACY
Proposed research	**2020**	**Support vector machine (SVM)**	**88%**
		Decision Tree (DT)	**84%**
		Neural Network (NN)	**77%**
[3]	2018	Naïve Bayes	76%
		SVM (Support vector machine)	65%
		Decision Tree	73%
[7]	2018	Decision Tree	75%

		Naive Bayes	71%
		KNN (_nearest neighbor)	65%
[8]	2018	SVM (Support vector machine)	91%
		Naïve Bayes	88
		KNN (K-nearest neighbor)	89%
[9]	2015	Decision tree (J48)	74.87%
		Naïve bayes	76.96%
[14]	2019	SVM (Support vector machine)	78%
		RF (Random Forest)	75%
		Naïve Bayes	73%
[15]	2016	J48	67%
		CART	62%
		SVM (Support vector machine)	65%
		KNN (K-nearest neighbor)	53%
[17]	2013	Logistic regression	93%
		Linear discriminant analysis	92%
		Fuzzy c-mean	86%
		SVM (Support vector machine)	98%
		NN (Neural Network)	93%
		RF (Random Forest)	93%
[21]	2015	J48	73.82%
		KNN (K-nearest neighbor)	K=1 70.81% K=2 72.65 K=5 73.17
		SVM (Support vector machine)	73.34%
		Random Forest	71.74
[24]	2015	Naïve bayes	77.8%
		C4.5	78%
		SVM (Support vector machine)	77.4%
		KNN (K-Nearest neighbor)	77.7%
[29]	2012	SVM (Support vector machine)	74.8%
		PNN	67%
		BLR	75%
		MLR	75%
[30]	2017	Naive bayes	90%
		RF (Random forest)	96%

		C4.5	100%
		Logistic regression	0.99%
[34]	2013	SVM (Support vector machine)	78%
[35]	2010	SVM (Support vector machine)	83.5%
[36]	2014	SVM (Support vector machine)	98%
[42]	2014	C4.5	86%
		SVM (Support vector machine)	74.8%
		K-NN	78%
		PNN	67%
		BLR	75%
[43]	2012	Bagging	77.4%
		AdaBoost	77.2%
		Random Forest	73.5%
		Multiclass	77.2%
[54]	2008	NN (Neural Network)	80%
		ANFIS	80.11%
[55]	2019	Decision Tree	77%
		Weighted KNN	77.3%
		Logistic Regression	79.3%
		Ensemble Method	80.60%
[56]	2017	Random forest	94%
		Decision tree	88%
		Naïve Bayes	91%
[65]	2017	Recurrent Deep Neural Network (RDNN)	Type 1 diabetes=78% Type 2 diabetes=81%
[73]	2016	Hybrid Model	96.09%
		Support vector machine (SVM)	75.22%
		Neural network (NN)	76.58%
[76]	2019	Logistic regression	Accuracy=0.7761
		Naïve bayes	Accuracy=0.7664
		Support vector machine (SVM)	Accuracy=0.7568
		KNN (K-nearest neighbor)	Accuracy=0.751
		Decision tree	Accuracy=0.6757
[77]	2008	ID3	80%
		Decision Tree	72%
[79]	2018	SVM (Support vector machine)	88.8%

		Naïve Net	88.54%
		Decision Stump	83.72%
		Ensemble method	90.36%

4.1 Summary

In this chapter, we have used two diabetes datasets in the R programming language tool for the prediction of diabetes disease. First, applied Random Forest (RF) as a feature selection method, and then applied different classification methods on both datasets. The datasets are divided into testing and training datasets with a ratio of 20% and 80% respectively. In the end, conclude the results of classification methods. The result shows that the Support Vector Machine (SVM) provides higher accuracy of early diagnosis of diabetes disease. In this chapter, all the results show in the form of tables and graphs. In the next chapter, the discussions about the results define and also explain the future work of this study.

V. DISCUSSION

The result shows that SVM is the best classifier for the classification of diabetes disease. This study also concluded that proposed research is helpful for the biologist in the early prediction of diabetes disease. This model is time and cost-effective for everyone and can also be used for diagnosing other diseases. Random forest is used to select important features that describing diabetes disease. Then, this study applies classification methods to classify diabetic and non-diabetic patients. Proposed research used an accuracy graph and confusion matrix to show the performance of the model and also used the histograms to represent the number of patients having diabetes. Future work is that this proposed model will be tested on a gene expression dataset having more than 10k or 50k features. The feature extraction method is useful to handle this type of huge data. By reducing features, this model will be very helpful and useful to diagnose diabetes disease in a short time and effective manner, it will be cost-effective also. In the future, this proposed model will be tested for the prediction of any type of disease.

5.1 Summary

This chapter explains the experimental results of the proposed study. It defines that the Random forest is the best feature selection method for reducing features. It also identifies the most suitable method on the basis of AU-ROC results. It gives the answers about those questions which are identified in the introduction of chapter I. Future work is also explained in this chapter. The next chapter concludes the result of the proposed study.

VI. CONCLUSION

Data mining plays a significant role in the diabetes dataset for the prediction of diabetes. Proposed research has tried to solve the problem of the prediction of diabetes disease. First of all, this study used Random Forest (RF) to extract the important features. After preprocessing the dataset, this research implements classification methods. The classification techniques used for predicting diabetes in patients are the Support vector machine (SVM), decision tree (DT), and neural network (NN) for the classification of diabetes patients by using two datasets. Then this study calculated the accuracy of these three classification methods on datasets without testing and training. After that, this study calculated Error rate and time in seconds of all. Then this research calculated accuracy with 20%

testing and 80% training and found the SVM shows higher accuracy.

6.1 Summary

In this chapter, the result shows that the random forest is the best feature selections method and the support vector machine provide the best classification results when data is partitioned as 80% training and 20% testing.

REFERENCES

[2] Wu, H., Yang, S., Huang, Z., He, J., & Wang, X. (2018). Type 2 diabetes mellitus prediction model based on datamining. Informatics in Medicine Unlocked, 10, 100-107.

[3] Sisodia, D., &Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms.Procedia ComputerScience, 132, 1578-1585.

[5] Osman, A. H., & Aljahdali, H. M. (2017). Diabetes disease diagnosis method based on feature extraction using K-SVM. *Int J Adv Comput Sci Appl*, *8*(1).

[7] Amina Azrar, Yasir Ali, Muhammad Awais and Khurram Zaheer, "Data Mining Models Comparison for Diabetes Prediction" International Journal of Advanced Computer Science and Applications(IJACSA), 9(8), 2018.

[9] Iyer, A., Jeyalatha, S., & Sumbaly, R. (2015). Diagnosis of diabetes using classification mining techniques. *arXiv preprint arXiv:1502.03774*.

[10] Joshi, R., & Alehegn, M. (2017). Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach. International Research Journal of Engineering and Technology, 4(10).

[13] Hashi, E. K., Zaman, M. S. U., & Hasan, M. R. (2017, February). An expert clinical decision support system to predict disease using classification techniques. In *2017 International Conference on Electrical, Computer and Communication Engineering (ECCE)* (pp. 396-400). IEEE.

[14] Sneha, N., & Gangil, T. (2019). Analysis of diabetes mellitus for early prediction using optimal features selection. *Journal of Big Data*, *6*(1), 13.

[16] Rajesh, K., & Sangeetha, V. (2012). Application of data mining methods and techniques for diabetes diagnosis. *International Journal of Engineering and Innovative Technology (IJEIT)*, *2*(3).

[20] Shukla, N., & Arora, M. (2016). Prediction of diabetes using neural network & random forest tree. *International Journal of Computer Sciences and Engineering*, *4*, 101-104.

[21] Kandhasamy, J. P., & Balamurali, S. (2015). Performance analysis of classifier models to predict diabetes mellitus. *Procedia Computer Science*, *47*, 45-51.

[23] kumar Dewangan, A., & Agrawal, P. (2015). Classification of diabetes mellitus using machine learning techniques. *International Journal of Engineering and Applied Sciences*, *2*(5).

[24] Thirumal, P. C., & Nagarajan, N. (2015). Utilization of data mining techniques for diagnosis of diabetes mellitus-a case study. *ARPN Journal of Engineering and Applied Science*, *10*(1), 8-13.

[25] Patil, R. N., & Tamane, S. C. (2018). Upgrading the performance of KNN and naïve bayes in diabetes detection with genetic algorithm for feature selection. *International Journal of Scientific Research in Computer Science*, *3*(1), 1371-1381.

[28] Devi, M. R., & Shyla, J. M. (2016). Analysis of various data mining techniques to predict diabetes mellitus. *International journal of applied engineering research*, *11*(1), 727-730.

[29] Karthikeyani, V., Begum, I. P., Tajudin, K., & Begam, I. S. (2012). Comparative of data mining classification algorithm (CDMCA) in diabetes disease prediction. *International Journal of Computer Applications*, *60*(12).

[31] Srikanth, P., & Deverapalli, D. (2016, February). A critical study of classification algorithms using diabetes diagnosis. In *2016 IEEE 6th International Conference on Advanced Computing (IACC)* (pp. 245-249). IEEE.

[39] Benbelkacem, S., & Atmani, B. (2019, April). Random Forests for Diabetes Diagnosis. In *2019 International*

ISSN: 2708-7123 | Volume-**01**, Issue Number-**03** | September-2020
LC INTERNATIONAL JOURNAL OF STEM
Web: www.logicalcreations.org/stem | www.lcjstem.com | DOI: https://doi.org/10.47150

Conference on Computer and Information Sciences (ICCIS) (pp. 1-4). IEEE.

[55] Dezfuli, S. A. M., Dezfuli, S. R. M., Dezfuli, S. V. M., & Kiani, Y. (2019). Early Diagnosis of Diabetes Mellitus Using Data Mining and Classification Techniques. *Jundishapur Journal of Chronic Disease Care*, *8*(3).

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

[57] Mercaldo, F., Nardone, V., & Santone, A. (2017). Diabetes mellitus affected patients classification and diagnosis through machine learning techniques. *Procedia computer science*, *112*, 2519-2528.s

[76] Choudhury, A., & Gupta, D. (2019). A survey on medical diagnosis of diabetes using machine learning techniques. In *Recent Developments in Machine Learning and Data Analytics* (pp. 67-78). Springer, Singapore.

[78] Kalantari, A., Kamsin, A., Shamshirband, S., Gani, A., Alinejad-Rokny, H., & Chronopoulos, A. T. (2018). Computational intelligence approaches for classification of medical data: State-of-the-art, future challenges and research directions. *Neurocomputing*, *276*, 2-22.

[79] Alehegn, M., Joshi, R., & Mulay, P. (2018). Analysis and prediction of diabetes mellitus using machine learning algorithm. *International Journal of Pure and Applied Mathematics*, *118*(9), 871-878.

[81] Kumar, N. K., Vigneswari, D., Krishna, M. V., & Reddy, G. P. (2019). An optimized random forest classifier for diabetes mellitus. In *Emerging Technologies in Data Mining and Information Security* (pp. 765-773). Springer, Singapore.