ISSN: 2708-7123 | Volume-**01**, Issue Number-**04** | December-2020
LC INTERNATIONAL JOURNAL OF STEM
Web:www.logicalcreations.org/stem | www.lcjstem.com | DOI: https://doi.org/10.47150

A COMPARATIVE ANALYSIS ON DIABETES DATASETS USING DATA MINING TECHNIQUES

Zeenat Bashir[1], Dr. Hamid Ghous[2]
Institute of Southern Punjab (ISP), Multan, Pakistan
[1]Zeenatbashir16@gmail.com[2]hamidghous@isp.edu.pk

ABSTRACT—In past, many researchers have been worked on the early diagnosis of diabetes disease. They used different diabetes datasets for the prediction of diabetes disease. Diabetes disease is one of the chronic diseases and becoming a cause of death among peoples. So many factors involved which cause diabetes disease and in this way a huge amount of data increasing about diabetes disease.In this paper, we compared different cases of diabetes disease based on two different Pima Indian diabetes datasets. From previous studies we found that different deep learning and machine learning methods had been applied to these two datasets but achieved very efficient methods used to diagnose diabetes. Also, determine the research gap in the application of different methods in the biomedical field.

Keywords—Diabetes disease, machine learning, deep learning, data mining, classification

I. INTRODUCTION

1.1 Background

Diabetes Disease:

In [9] 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, ShujaMirza& Dr. Sonumittal (2018), AiswaryaIyer et al. (2015).

In [3], [41] diabetes is thecause 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, RaminGhorbania&RouzbehGhousi**(**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, DeeptiSisodia&Dilip Singh Sisodia(2018) [3].

In [42] 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).

In [43] 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

ISSN: 2708-7123 | Volume-01, Issue Number-04 | December-2020
LC INTERNATIONAL JOURNAL OF STEM
Web:www.logicalcreations.org/stem | www.lcjstem.com | DOI: https://doi.org/10.47150

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.

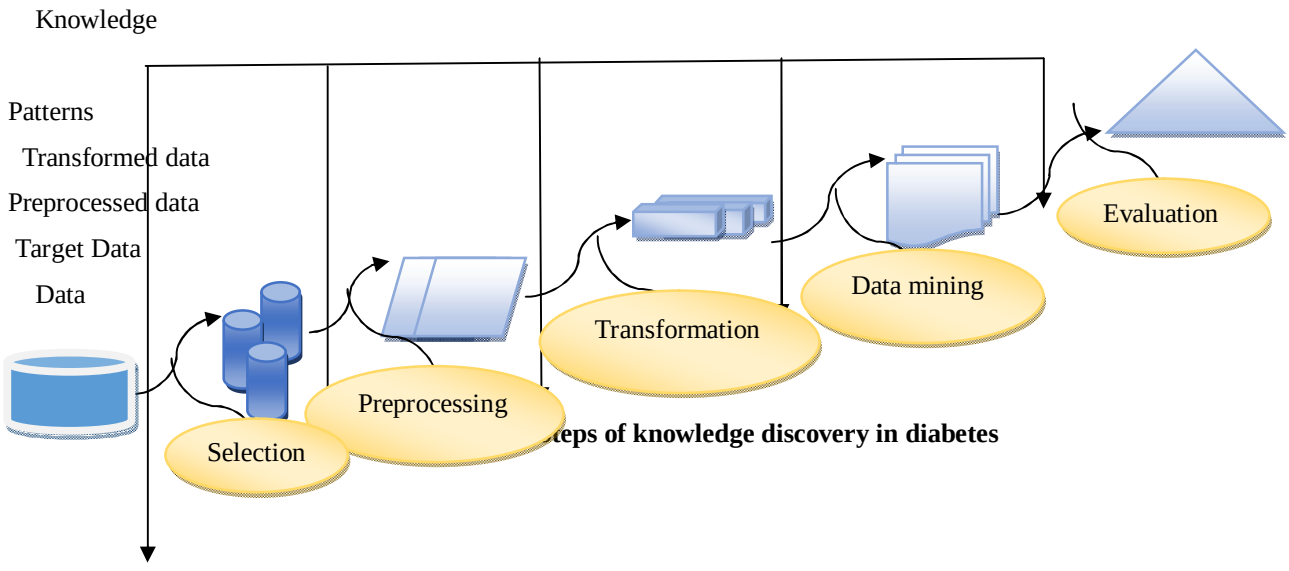
3. **Gestational diabetes:**

In [43] 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).

In [9] 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, AiswaryaIyer et al., (2015). In [9] 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 thecomplications related to them, if these are detected in the early stages, AiswaryaIyer et al., (2015).

Knowledge discovery in databases (KDD)

In [44] 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, RaminGhorbani&RouzbehGhousi, (2019):



teps of knowledge discovery in diabetes

1.2 Data Mining

In [44] 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, RaminGhorbani&RouzbehGhousi (2019).

In [42] 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).

In [43] 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).

In [9] 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 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, AiswaryaIyer et al., (2015).

1.3 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):

In [3] 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, DeeptiSisodia&Dilip Singh Sisodia (2018).

II. Decision Tree:

In [3] 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

ISSN: 2708-7123 | Volume-01, Issue Number-04 | December-2020
LC INTERNATIONAL JOURNAL OF STEM
Web:www.logicalcreations.org/stem | www.lcjstem.com | DOI: https://doi.org/10.47150

attributes, decision trees select every node for obtaining the highest information, DeeptiSisodia&Dilip Singh Sisodia, (2018).

III. Neural Network:

In [45] 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).

1.4 Feature Selection

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

In [46] 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).

1.5 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.

II. LITERATURE REVIEW

In past many researcher have been worked on early diagnosis of diabetes disease. As diabetes disease is one of the chronic disease and becoming a cause death among peoples. So many factors involves 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.

Background section is divided into sub-section; Section 1 describes the study of previous work using PIMA Indian diabetes dataset 1 (UCI Repository dataset), Section 2 describes the previous study on PIMA Indian diabetes dataset 2 (kaggle dataset). Table 1 shows the comparative study of previous research work.

2.1 PIMA Indian diabetes Dataset 1 (UCI Repository dataset)

In this section, the previous work done in the field of diabetes disease diagnosis using PIMA Indian diabetes dataset is discussed.

Rahul C. (2015) Machine learning algorithms are mostly used in the medical field for diagnosing different diseases. For obtaining higher accuracy of prediction diabetes disease many researchers used machine learning (ML) algorithms. Different machine learning categories are used such as supervised machine learning algorithms and unsupervised machine algorithms for predicting diabetes [1].

DeeptiSisodia&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 [3].

Vosoulipour et al., (2008) used an artificial neural network and Adaptive network-based interface system (ANIFS) for the classification of diabetes disease. Pima Indian diabetes dataset used which is taken from the UCI repository to classify the diabetic or non-diabetic patients into two classes. The dataset consists of eight attributes, 768 instances, and two classes. These two classes define diabetic or non-diabetic patients. Machine learning methods used this dataset for diagnosing diabetes. Genetic algorithms are used as an important feature selection in this study. After normalizing features they calculate the performance of the network based on the training and testing dataset. They used four features that are collected from genetic algorithms, as input into ANFIS, and obtain the best results. After comparison, ANFIS provides the best results and better performance than an artificial neural network [4].

Rahul Joshi&MinyechilAlehegn (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 [10].

N. Sneha1 &TarunGangil (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% [14].

Raj Anand et al., (2013) utilized the K-fold cross-validation method to classify diabetes disease. The Pima Indian dataset was utilized for this investigation which is collected from the UCI machine learning repository. The dataset consists of nine attributes and 768 instances. With Principle component analysis (PCA) data preprocessing required for improvement of predictive accuracy. Performance of classification accuracy evaluated with Principle Component Analysis (PCA) preprocessing and High Order Neural Network (HONN). The results obtained that the PCA preprocessing used for minimizing the mean square error and HONN provide higher order classification of diabetes [19].

Neha Shukla &Meena Arora (2016) used the PIMA Indian dataset for the prediction of diabetes. The aim of the proposed research is to predict the diabetes disease done through some classification algorithms. A scaled conjugate gradient (Neural Network) technique is used to extract useful information about diabetic patients. The performance was evaluated on the base of sensitivity and specificity. Classification methods, neural network, and Random forest tree used for prediction of diabetes. The results demonstrate that the random forest tree gives better precision of 92.96% than the neural network [20].

J.PradeepKandhasamy&S.Balamurali (2015) focused on predicting diabetes using data mining techniques. They also compare the performance of all algorithms which are used to diagnose the diabetes disease. Machine learning classifiers such as the J48 decision tree, k-nearest neighbors, and random forest and support vector machines were used to classify the diabetes patients. Performances of algorithms measured through the University of California Irvine (UCI) machine

learning repository. The researchers did not describe the preprocessing which is applied to the dataset; they just explain that the noise was removed from the data. The performance of all algorithms evaluated in terms of sensitivity, specificity, and accuracy. The results are calculated with two different situations in which one is before pre-processing and the other is after pre-processing. In the first situation, the J48 classifier provides higher accuracy of 73.82% than other classifiers. In other situations, the dataset gives accurate accuracy; in this case, KNN and random forest provide 100% accuracy than other all classifiers [21].

M. Durairaj&G.Kalaiselvi (2015) focused on the Backpropagation network for the diagnosis of diabetes disease. Layered feed-forward artificial neural network ANNs used for backpropagation. It reduces the error for enhancing the performance of the diagnosis of diabetes. Pima Indian diabetes dataset used for this experiment which is collected from the UCI machine learning repository and it consists of nine attributes and 768 instances. The experiment was evaluated on the basis of training and testing of the dataset. In this study, the Levenberg Marquardt (LM) algorithm is used for backpropagation which gives higher performance results than previous researches algorithms. The proposed research provides higher accuracy of 91% with a backpropagation network of LM than previous researches [22].

Amit Kumar Dewangan&Pragati Agarwal (2015) presented a diagnosis of diabetes using data mining techniques. They collected the PIMA Indian diabetes dataset from the UCI Repository which is used in this technique. The work is divided into three stages. In the first stage, the accuracy of classification is achieved with the individual model. In the second stage, a hybrid model is developed for obtaining higher accuracy. In the third and last stage, the feature selection method is implemented on the best hybrid model for obtaining higher accuracy. The prediction of diabetes is done through the artificial neural network, K-fold cross-validation and classification, support vector machine, K-nearest neighbor method, and data mining algorithms. The result shows that the new hybrid model gives better accuracy in terms of accuracy, specificity, and sensitivity [23].

Thirumal P. C. &Nagarajan N. (2015) focused on data mining techniques for diagnosing diabetes. The main objective of this research is to provide the best algorithm for the prediction of diabetes. Naïve Bayes classifier, C4.5 algorithm, KNN algorithm, and SVM used for an early stage diabetes diagnosis. These four algorithms are significant for automatic classification tools which is helpful for the researcher to reduce the processing time. The dataset PIMA Indian diabetes is used to measure the performance of algorithms which is available at the UCI machine learning repository. C4.5 gives a higher accuracy of 78.25% and KNN provides a minimum accuracy of 77.73% than other algorithms [24].

Ratna Nitin Patil&Dr.Sharvari Chandra shekharTamane (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 [25].

Md. Maniruzzaman et al., (2017) used four classification methods with three different kernels for classifying diabetes disease. Classification methods are Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Naïve Bayes (NB), and Gaussian Process Classification (GPC). In this study, the Gaussian process classification technique used three kernels such as the linear kernel, Polynomial kernel, and Radial base kernel for obtaining accuracy. Pima Indian diabetes dataset used for this experiment which is collected

from the UCI repository consists of nine attributes and 768 instances. Six performance factors utilized for performance evaluation which are Sensitivity (SE), Accuracy (ACC), Specificity (SP), Positive Predictive Value (PPV), Negative Predictive Value (NPV), and Receiver Operating Curve (ROC). It is concluded that on the basis of all performance factors, GP based classification technique with RBF kernel provides higher accuracy than other all algorithms [26].

NirmalaDevi.M et al., (2013) proposed a hybrid model for increasing the accuracy rate of predicting diabetes. For the classification of diabetes, researchers used the Pima Indian diabetes database. In the preprocessing step it combines the k-Nearest Neighbor (KNN) with k-means and reduces the noisy data, also replaced missing values. The objective of this study is to improve classification accuracy with the help of the proposed hybrid model. Tenfold cross-validation utilizes for performance enhancement. Performance is evaluated on the basis of accuracy, sensitivity, and specificity. The proposed model compared with simple K-Nearest Neighbor (KNN) and K-means K-Nearest Neighbor and the result showed that with multi-step preprocessing it provides higher accuracy than single classification methods. This amalgam KNN obtain an accuracy of 97.4%, simple KNN accuracy of 73.17%, and K-means and KNN accuracy of 97% [27].

Dr. M. Renuka Devi & J. Maria Shyla (2016) implemented many data mining techniques for early diabetes detection. PIMA Indian diabetes dataset was used to obtain the accuracy of data mining techniques in diagnosing disease. This dataset collected from UCI Repository and it consists of 768 instances which are used to obtain the higher accuracy of classification in prediction. Naïve Bayes, MLP, Bayesian network, C4.5, Amalgam KNN, ANFIS, PLS-LDA, J48, GA, Generic GA, ANN, Homogeneity, SVM, BLR are analyzed to predict the diabetes disease. THE modified J48 classifier provides higher accuracy of 99.87% than others [28].

Han Wu et al., (2017) proposed a model which is based on data mining techniques for diagnosing type2 diabetes mellitus. They developed a hybrid model of K-Means and the Logistic Regression algorithm to achieve a high prediction accuracy rate. The Pima Indians Diabetes Dataset and the Waikato Environment for Knowledge Analysis toolkit were used for diagnosing diabetes. This model ensures the high quality of experimental data. It provides less time consuming and maximum retention of original data. The results show that the proposed model achieved higher prediction accuracy [29].

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 [31].

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 [33].

V. AnujaKumari& R. Chitra (2013) presented a classification method for diagnosing diabetes with less time and better performance. They used Pima Indian diabetes disease dataset which is collected from the UCI Repository for diabetes classification. The proposed method Support vector machine algorithm classifies the diabetic or non-diabetic patients. This supervised machine learning method classifies diabetes disease from a large dataset. Performance is evaluated on the base of accuracy, sensitivity, and specificity. SVM provides an accuracy of 78% with a sensitivity of 80% and specificity of 76.5% [34].

MinyechilAlehegn 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% [35].

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 [38].

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 [39].

2.2 PIMA Indian diabetes dataset 2 (kaggle dataset)

In this section, the previous work done in the field of diabetes disease diagnosing using PIMA Indian diabetes dataset 2 (kaggle dataset) is discussed.

SeyedAtaaldinMahmoudinejad 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 [2].

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 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% [5].

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 [6].

AnjliNegi&VarunJaiswal (2016) used Support Vector Machine (SVM) to predict diabetes disease. The Pima Indian Diabetes Dataset is used which contains 49 attributes and 102,538 instances. In this dataset, 64,419 tested positive, and the remaining 38,115 tested negative. In this research, the researchers did not explain the use of attributes. For replacing the missing values first of all preprocessed the dataset and also normalization method applied between 0 and 1 values. Before the classification of diabetes, they used different feature selection methods for extracting important features. Fselect script selected four attributes from LIBSVM and Wrapper and Ranker methods selected nine and twenty attributes from the Weka toolkit. For the evaluation of performance, researchers used a 10-cross fold validation technique. After combined all dataset, it provides higher accuracy of 72% and the prediction of diabetes disease become more reliable than traditional researches [7].

Ebenezer ObaloluwaOlaniyi&Kashman Adnan (2014) proposed a multilayer feed forwarding network for

improvement of diabetes prediction accuracy. For training the algorithms they used the backpropagation algorithm. Before preprocessing, researchers normalized the dataset for obtaining higher accuracy of classification. Pima Indian diabetes dataset was used in this study in which 500 instances for training test and 268 instances for the testing set. The result showed that the accuracy obtains 82% which is higher than previous studies. This proposed accuracy compared with other previous researches of diabetes where different types of algorithms used such as the C4.5 algorithm, the nearest neighbor with feature selection, ADAP, and EM algorithm. The accuracy achieved from multilayer feed forwarding is higher than all other algorithms [8].

AiswaryaIyeret al., (2015) proposed solutions for diagnosing diabetes disease at an early stage using classification methods such as the J48 decision tree and naïve Bayes. Data tested by cross-validation technique and percentage split technique. For pre-paring training and test data 10-cross validation used. After pre-processing, data divided into tested negative or tested positive on the base of final results. The classification types of data mining are applied to the Pima Indians Diabetes Database of National Institute of Diabetes and Digestive and Kidney Diseases. The result is that both models are efficient in diagnosing diabetes with higher accuracy [9].

ZahedSoltani& Ahmad Jafarian (2016) used a probabilistic artificial neural network for diagnosing type II diabetes disease. This PANN method was implemented on MATLAB and Pima Indian diabetes dataset used which consists of 768 instances and it was collected from the UCI Repository. Performance of experiment measures on the basis of 90% of the training set and 10% of the testing dataset. A preprocessing technique did not use in this research. The result showed that the training and testing accuracy of 89.56% and 81.49% which is good than other all researches [11].

SomnathRakshit et al., (2017) proposed a two-class predicting type II diabetes mellitus. In this study, Pima Indian diabetes dataset was utilized in which seven important factors were selected for the experiment. This experiment was done on the basis of an 80 % training set and a 20% testing set of female patients with 21 years old. For the normalization of data samples, researchers first of all preprocessed the data through mean and standard deviation for obtaining numerical values. Using the correlation method they extract important features. The result showed that a higher accuracy of the proposed model of 83.3% [12].

EmranaKabirHashi 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 experimental results. The result shows that the decision tree gives better accuracy of 90.43% for diagnosing diabetes disease [13].

MammanMamuda&SarathaSathasivam (2017) implemented three supervised machine learning algorithms for the prediction of survivals of diabetes disease. Scaled Conjugate Gradient (SCG), Bayesian Regulation (BR), and Levenbreg Marquardt (LM) are three machine learning algorithms used in this study. They used Pima Indian diabetes dataset with eight attributes and 768 instances for performance evaluation with the help of MATLAB. 10-cross validation technique used with training and testing dataset. The researcher evaluated the result of the proposed study, Levenberg Marquardt provides higher accuracy of predicting diabetes disease then Bayesian regulation and scaled conjugate gradient. The result was evaluated on the basis of Mean Squared Error (MSE) of 0.00025019 which is good than previous results [15].

K. Rajesh & V. Sangeetha (2012) focused on the classification of diabetes with the help of the data mining technique. They also predict the patient being affected by diabetes or not.

Different classification techniques are applied for diagnosing diabetes. Pima Indian diabetes dataset used for data mining classification which is collected from the UCI Repository contains 768 instances with 8 attributes. All algorithms provide different accuracy rate for diagnosing diabetes but C4.5 is a common decision tree which provides higher accuracy of 91% from other classification algorithms [16].

AkmAshiquzzaman et al., (2017) used Multilayer Perceptron Deep Learning Neural Network, General Regression Neural Network (GRNN), and Radial Basis Function (RBF) for prediction of diabetes. In this experiment Pima Indian diabetes dataset used which is collected from the UCI Repository is held by the National Institute of Diabetes and Digestive and Kidney Diseases. This dataset consists of eight attributes and 768 instances, female patients are tested in this experiment with the age of 20 years. The performance was evaluated on the basis of 576 training samples and 192 testing samples. In this study, the result showed that the accuracy achieved 88.41% through deep learning neural networks [17].

Sushant Ramesh et al., (2017) applied deep learning neural network for diabetes prediction. For the prediction of two types of diabetes, disease researchers used Recurrent Neural Network. They utilized Pima Indian diabetes dataset which is collected from the UCI machine learning repository. It consists of 768 instances and eight attributes. First of all, implemented a feature selection method for extracting high priority features such as Glucose, BMI (Mass), Age, Pregnancies, Pedigree function, Blood pressure, skin thickness, and insulin. The performance of the experiment was evaluated on the basis of 80% training set and 20% testing dataset. The result showed that the accuracy of type II is higher than type I. Type II provides an accuracy of 81% and type I accuracy of 78% [18].

NASIB SINGH GILL & POOJA MITTAL (2016) focused on developing a hybrid prediction model for predicting diabetes disease using MATLAB tool. From the UCI machine learning repository, Pima Indian diabetes dataset was collected for this experiment. This dataset consists of nine attributes and 768 instances. First of all applied filtration method applied for feature selection to reduce the redundancy of data. In second stage classification methods, the Support vector machine (SVM) and Neural network (NN) are used for prediction of accuracy. Performance is evaluated on the basis of recognition rate, Mean Absolute Error (MAE), and Receiver Operation Curve (ROC). The result showed that this proposed model provides higher accuracy of 96.09% after comparing it with the previous researches model [30].

Jianchao Han et al., (2008) focused on developing a novel model using the Rapid Miner tool for the prediction of diabetes disease. Pima Indian diabetes dataset was used, which is collected from the UCI repository. Dataset consists of 768 instances with eight attributes of female patients at least 21 years old. Preprocessing of data is required on the basis of feature extraction and selection, missing data removal, normalization of data, and discovery of hidden data relationship, analysis of visual data, and at the end construct the prediction model. Decision tree and ID3 classification methods are used to improve performance accuracy. The result showed that ID3 provides higher accuracy of 80% than the decision tree which provides an accuracy of 72% [32].

Khyati K. Gandhi&Prof.Nilesh B. Prajapati (2014) enhanced the efficiency, predictive accuracy, and reduced the complexity of classification results. They used two feature selection techniques for selecting significant features from the Pima Indian diabetes dataset which is selected from the National Institute of Diabetes and Digestive and Kidney Disease. After pre-processing of the dataset, informative features were selected with two feature selection techniques such as F-score and K-means clustering. Then SVM classifier performance was evaluated on the basis of accuracy, sensitivity, specificity, and AUC. The proposed support vector machine improved and provides an accuracy of 98% with a sensitivity of 97.77%, the specificity of 97.79, and AUC 0.9 [36].

Asha GowdaKaregowda et al., (2012) focused on developing a new hybrid model for the classification of diabetes disease using K-means and K-Nearest Neighbor. Pima Indian diabetes dataset is used for an experiment that is collected from the UCI machine learning repository. This dataset has two categories of tested positive and tested negative; both consist of eight attributes and 768 instances. Preprocessing is necessary for the dataset for enhancing classification accuracy. The performance was evaluated on the basis of 70% training and 30% testing dataset. In this study proposed model consists of three stages, in first stage instances identification and reducing irrelevant samples used k-means clustering. Genetic algorithm (GA) and Correlation-based feature selection (CFS) are used in the second stage for the extraction of useful features. In the third and last stage, classification is done through K-nearest neighbor after using the first and second stages. Result improved in classification using KNN and cascaded K-means along with GA-CFS. The proposed model achieved higher accuracy of 96.68% for the prediction of diabetes compared to previous researches [37].

Muhammad WaqarAslam et al., (2013) used the Pima Indian diabetes dataset and applied some genetic programming based methods to build a model for the prediction of diabetes disease. Diabetes disease is increasing rapidly all over the world. With prior knowledge, Genetic programming is used to generate new features with a combination of existing diabetes features. In the proposed research they used three stages; the first stage is feature selection in which used different tests such as the t-test, Kolmogorov-Smirnov test, kullback-Leibler divergence test, F-score selection, and GP. In the second stage, genetic programming is used to generate new features with a comparison of original features. The performance of classification is tested with a k-nearest neighbor and support vector machine in the last stage. The genetic programming-support vector machine provides higher accuracy [40].

2.3 Comparison of previous researchers using diabetes datasets

Below is the comparison table between different previous researchers using diabetes datasets. They used different techniques on the diabetes disease dataset in their research. They implemented different methods in different ways on the diabetes dataset and to classify the patients having diabetes disease or not. The comparison of accuracy and methods which are used by previous researchers are given in table 1.

ISSN: 2708-7123 | Volume-01, Issue Number-04 | December-2020
LC INTERNATIONAL JOURNAL OF STEM
Web:www.logicalcreations.org/stem | www.lcjstem.com | DOI: https://doi.org/10.47150

Sr.#	Ref #	Year	Dataset	Methodology	Objective	Achievements
1	[2]	2019	PIMA Indian diabetes dataset	Decision Tree, Weighted KNN, Logistic Regression, Ensemble Method	Prediction of diabetes	Ensemble method provide higher accuracy of 80.60%
2	[3]	2018	Pima Indians Diabetes Dataset	Decision Tree, support vector machine (SVM) and Naive Bayes	detection of diabetes at its early stage	Naive Bayes classification algorithm provide higher accuracy
3	[4]	2008	Pima Indian diabetes dataset	Neural network (NN), ANFIS	Prediction of diabetes	ANFIS provide higher accuracy of 80.11%
4	[5]	2017	Pima Indian diabetes dataset	Random forest, Decision Tree, Naïve Bayes	Diagnosis of diabetes	Random forest provide higher accuracy of 94%
5	[6]	2017	Pima Indian diabetes dataset	J48, MLPNN, Hoeffding Tree, JRip, BayesNet, Random Forest	Prediction of diabetes	Different methods provide higher accuracy of Precision=0.76 Recall = 0.76 F-Measures= 0.76
6	[7]	2016	Pima Indian diabetes dataset	Support Vector Machine (SVM)	Prediction of diabetes	Novel model provide higher accuracy 72.93%
7	[8]	2016	Pima Indian diabetes dataset	Support vector machine (SVM)	Prediction of diabetes	Novel model provide higher accuracy 72.93%
8	[9]	2015	Pima Indians Diabetes Database	Decision Tree and Naïve Bayes algorithm	diabetes prediction	diagnosis of diabetes for local and systematic treatment
9	[10]	2017	Pima Indian Diabetes Data set	KNN, Naïve Bayes, Random forest, and J48	Analysis and prediction of diabetes diseases	decision tree provided high accuracy for predicting diabetes disease
10	[11]	2016	Pima Indian diabetes dataset	Probablistic Neural Network (PNN)	Diabetes prediction	Training accuracy=89% Testing accuracy=81%
11	[12]	2017	Pima Indian diabetes dataset	Two class neural network	Classification of diabetes	Two class neural network provide accuracy 83.3%
12	[13]	2017	Pima Indians diabetes data set	Decision Tree and K-Nearest Neighbor (KNN)	Prediction of diabetes	Decision tree provided best accuracy for diagnosing diabetes
13	[14]	2019	dataset is collected from UCI machine repository	Decision tree, Random forest, KNN, SVM, NB	Diagnosis early stage of diabetes mellitus	Decision tree and Random forest algorithm provide higher accuracy for diagnosing diabetes.
14	[15]	2017	Pima diabetes disease dataset	The Levenberg Marquardt Learning Algorithm (LMLA), Bayesian Regulation Learning	Diabetes prediction	Mean Square Error (MSE)= 0.00025019 Mean Square Error

ISSN: 2708-7123 | Volume-01, Issue Number-04 | December-2020
LC INTERNATIONAL JOURNAL OF STEM
Web:www.logicalcreations.org/stem | www.lcjstem.com | DOI: https://doi.org/10.47150

				Algorithm (BRLA), Scaled Conjugate Gradient Learning Algorithm (SCGLA)		(MSE)= 2.021e-05 Mean Square Error (MSE)= 8.3583
15	[16]	2012	Pima Indians diabetes data set	C-RT, CS-RT ID3, K-NN, LDA, NAÏVE BAYES, PLS-DA, SVM, RND TREE	Diagnosing diabetes	C 4.5 provide higher accuracy for diagnosing diabetes
16	[17]	2017	Pima Indian diabetes dataset	Deep Learning Architecture(MLP/GRNN/RBF) (General Regression neural network/Radial basis function)	Prediction of diabetes	Deep learning provide accuracy of 88.41%
17	[18]	2017	Pima Indian diabetes dataset	Recurrent Deep Neural Network (RDNN)	Classification of diabetes	Type II diabetes provide accuracy of 81%
18	[20]	2016	PIMA Indian dataset	Neural network and Random forest tree	Prediction of diabetes	Random Forest Tree give better precision result
19	[21]	2015	Pima Indian diabetic dataset UCI machine learning repository	J48 decision tree, KNN classifier, random forest, support vector machine	Classification of diabetes	J48 give higher accuracy than other three classifier
20	[22]	2015	Pima Indian diabetes dataset	Back propagation with Levenberg Marquardt (LM)	Prediction of diabetes	Back propagation with LM provide higher accuracy of 91% than previous researches
21	[23]	2015	PIMA Indian diabetes dataset	K-fold cross validation and classification, support vector machine, K-nearest neighbor method and data mining algorithms	Diagnosing diabetes	Hybrid model gives higher accuracy
22	[24]	2015	PIMA Indian diabetes dataset	Naïve bayes, decision tree, SVM and KNN	Diagnosis diabetes	Decision tree give the higher accuracy
23	[25]	2018	PIMA Indian diabetes dataset	K nearest neighbor and Naive Bayes used	Detection of diabetes	Proposed model give higher accuracy than earlier models
24	[26]	2017	Pima Indian diabetes dataset	LDA, QDA, NB and GPC	Classification of diabetes disease	GPC(Gaussian process classification) provide higher accuracy of 81.44%
25	[27]	2013	Pima Indian diabetes dataset	Simple KNN, K-means and KNN, Amalgam KNN	Prediction of diabetes disease	Amalgam KNN provide higher accuracy of 97.4%

ISSN: 2708-7123 | Volume-01, Issue Number-04 | December-2020
LC INTERNATIONAL JOURNAL OF STEM
Web:www.logicalcreations.org/stem | www.lcjstem.com | DOI: https://doi.org/10.47150

26	[28]	2016	PIMA Indian diabetes dataset	Naïve bayes, MLP, Byesian network, C4.5, Amalgam KNN, ANFIS, PLS-LDA, J48, GA, Generic GA, ANN, Homogenity, SVM, BLR	Predicting early stage diabetes disease	Modified J48 provide higher accuracy
27	[29]	2017	Pima Indians Diabetes Dataset	K-means algorithm and the logistic regression algorithm	predicting type 2 diabetes mellitus (T2DM)	enhancement of prediction accuracy
28	[30]	2016	Pima Indian diabetes dataset	Hybrid Prediction Model (HPM) using SVM+NN	Prediction of diabetes	Hybrid model provide higher accuracy of 96.09%
29	[35]	2018	Pima Indian diabetes dataset	SVM, Decision Stump, Naïve Net an Ensemble method	Prediction of diabetes	Proposed Ensemble Method provide higher accuracy (PEM)
30	[36]	2014	Pima Indian diabetes dataset	F-Score and K-means as feature selection, SVM as classification	Prediction of diabetes	Performance of SVM is improved than earlier researches
31	[37]	2012	Pima Indian diabetes dataset	Cascaded K-means clustering with K-Nearest Neighbor	Prediction of diabetes	Proposed model provide higher accuracy of 96.68%
32	[38]	2019	Diabetes dataset	Optimized Random Forest with Genetic Algorithm (GA-ORF)	Prediction of diabetes	Proposed hybrid model provides higher accuracy
33	[39]	2019	Pima Indian diabetes dataset	Random forest, C4.5, SVM, REPTree, BFTree	Prediction of diabetes	Random forest proved more efficient than other all algorithms
34	[40]	2013	Pima Indian diabetes dataset	GP-KNN, GP-SVM	Diagnosis diabetes	GP-SVM provide higher accuracy (82%)

Table 1:Comparison of previous researchers using diabetes datasets

At the above, comparative analysis of previous studies using Pima Indian diabetes datasets is given. The above table shows the previous study of different scholars. During the last decades, many researchers have been proposed hybrid and single classifier based models that were implemented on Pima Indian diabetes datasets. Some researchers used basic classifiers and some implemented these basic classifiers with the combination of different classifiers and different feature selection methods. Some researchers applied hybrid models on Pima Indian diabetes datasets to improve the performance of the prediction models.

III. CONCLUSION

This paper focuses on the comparison of data mining techniques on two diabetes disease datasets PIMA Indian diabetes UCI repository dataset and PIMA Indian diabetes Kaggle dataset. In this assessment paper, we reviewed the previous study that is conducted in past ten years. In this study, we analyzed that some data mining techniques are applied on UCI dataset and calculated their results then same techniques are applied on kaggle dataset but calculated different results. So, we conclude that same data mining techniques showed different results on these datasets.

Future work is that this proposed model can be implemented on a gene expression dataset that has 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 can also be used for the prediction of any type of disease.

REFRENCES

[1] Deo, R. C. (2015). Machine learning in medicine. *Circulation*, *132*(20), 1920-1930.

[2] 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).

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

[4] Vosoulipour, A., Teshnehlab, M., &Moghadam, H. A. (2008). Classification on diabetes mellitus data-set based-on artificial neural networks and ANFIS. In *4th Kuala Lumpur International Conference on Biomedical Engineering 2008* (pp. 27-30). Springer, Berlin, Heidelberg.

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

[6] 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

[7] Negi, A., &Jaiswal, V. (2016, December). A first attempt to develop a diabetes prediction method based on different global datasets. In *2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC)* (pp. 237-241). IEEE.

[8] Olaniyi, E. O., & Adnan, K. (2014). Onset diabetes diagnosis using artificial neural network. *Int. J. Sci. Eng. Res*, *5*(10), 754-759.

[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).

[11] Soltani, Z., &Jafarian, A. (2016). A new artificial neural networks approach for diagnosing diabetes disease type II. *International Journal of Advanced Computer Science and Applications*, *7*(6), 89-94.

[12] Rakshit, S., Manna, S., Biswas, S., Kundu, R., Gupta, P., Maitra, S., & Barman, S. (2017, March). Prediction of Diabetes Type-II Using a Two-Class Neural Network. In *International Conference on Computational Intelligence, Communications, and Business Analytics* (pp. 65-71). Springer, Singapore.

[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.

[15] Mamuda, M., &Sathasivam, S. (2017, August). Predicting the survival of diabetes using neural network. In *AIP Conference Proceedings* (Vol. 1870, No. 1, p. 040046). AIP Publishing LLC.

[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).

[17] Ashiquzzaman, A., Tushar, A. K., Islam, M. R., Shon, D., Im, K., Park, J. H., ...& Kim, J. (2018). Reduction of overfitting in diabetes prediction using deep learning neural network. In *IT Convergence and Security 2017* (pp. 35-43). Springer, Singapore.

[18] Ramesh, S., Balaji, H., Iyengar, N. C. S., &Caytiles, R. D. (2017). Optimal predictive analytics of pima diabetics using deep learning. *International Journal of Database Theory and Application*, *10*(9), 47-62.

[19] Anand, R., Kirar, V. P. S., & Burse, K. (2013). K-fold cross validation and classification accuracy of pima Indian diabetes data set using higher order neural network and PCA. *Int. J. Soft Comput. Eng*, *2*(6), 436-438.

[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.

[22] Durairaj, M. (2015). PREDICTION OF DIABETES USING BACK PROPAGATION ALGORITHM.

[23] kumarDewangan, 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.

[26] Maniruzzaman, M., Kumar, N., Abedin, M. M., Islam, M. S., Suri, H. S., El-Baz, A. S., &Suri, J. S. (2017). Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm. *Computer methods and programs in biomedicine*, *152*, 23-34.

[27] NirmalaDevi, M., alias Balamurugan, S. A., &Swathi, U. V. (2013, March). An amalgam KNN to predict diabetes mellitus. In *2013 IEEE International Conference ON Emerging Trends in Computing, Communication and Nanotechnology (ICECCN)* (pp. 691-695). IEEE.

[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] 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.

[30] Gill, N. S., & Mittal, P. (2016). A computational hybrid model with two level classification using SVM and neural network for predicting the diabetes disease. *J. Theor. Appl. Inf. Technol*, *87*(1), 1-10.

[31] 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.

[32] Han, J., Rodriguez, J. C., &Beheshti, M. (2008, December). Diabetes data analysis and prediction model discovery using rapidminer. In *2008 Second international conference on future generation communication and networking* (Vol. 3, pp. 96-99). IEEE.

[33] 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.

[34] Kumari, V. A., &Chitra, R. (2013). Classification of diabetes disease using support vector machine. *International Journal of Engineering Research and Applications*, *3*(2), 1797-1801.

[35] 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.

[36] Gandhi, K. K., &Prajapati, N. B. (2014). Diabetes prediction using feature selection and classification. *International journal of advance Engineering and Research Development*, *1*(05).

[37] Karegowda, A. G., Jayaram, M. A., &Manjunath, A. S. (2012). Cascading k-means clustering and k-nearest neighbor classifier for categorization of diabetic patients. *International Journal of Engineering and Advanced Technology*, *1*(3), 147-151.

[38] 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.

[39] Benbelkacem, S., &Atmani, B. (2019, April). Random Forests for Diabetes Diagnosis. In *2019 International Conference on Computer and Information Sciences (ICCIS)* (pp. 1-4). IEEE.

ISSN: 2708-7123 | Volume-**01**, Issue Number-**04** | December-2020
LC INTERNATIONAL JOURNAL OF STEM
Web:www.logicalcreations.org/stem | www.lcjstem.com | DOI: https://doi.org/10.47150

[40] 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.

[41] Ghorbani, R., &Ghousi, R. (2019). Predictive data mining approaches in medical diagnosis: A review of some diseases prediction. *International Journal of Data and Network Science*, *3*(2), 47-70.

[42] Das, H., Naik, B., &Behera, H. S. (2018). Classification of diabetes mellitus disease (DMD): a data mining (DM) approach. In *Progress in computing, analytics and networking* (pp. 539-549). Springer, Singapore.

[43] Azrar, A., Ali, Y., Awais, M., &Zaheer, K. (2018). Data mining models comparison for diabetes prediction. *Int. J. Adv. Comput. Sci. Appl.(IJACSA)*, *9*, 320-323.

[44] Ghorbani, R., &Ghousi, R. (2019). Predictive data mining approaches in medical diagnosis: A review of some diseases prediction. *International Journal of Data and Network Science*, *3*(2), 47-70.

[45] Fodor, I. K. (2002). *A survey of dimension reduction techniques* (No. UCRL-ID-148494). Lawrence Livermore National Lab., CA (US).

[46] Huang, Y., McCullagh, P., Black, N., & Harper, R. (2004, July). Feature selection and classification model construction on type 2 diabetic patient's data. In *Industrial Conference on Data Mining* (pp. 153-162). Springer, Berlin, Heidelberg.