ISSN 2723-6471

Machine Learning Techniques for Diabetes Prediction: A Comparative Analysis

Hoda A. Abdelhafez[1,*], Abeer A. Amer[2]

[1]*Information Technology Department, Princess Nourah bint Abdulrahman University, Riyadh, 11671, KSA*

[1]*Faculty of Computers and Informatics, Suez Canal University, Ismailia, Egypt*

[2]*Computer Science and Information Systems Department, Sadat Academy for Management and Sciences, Alexandria, 21525, Egypt*

(Received: February 20, 2024; Revised: April 1, 2024; Accepted: May 3, 2024; Available online: May 31, 2024)

Abstract

Diabetes mellitus, characterized by chronic hyperglycemia, presents significant challenges due to its associated complications and increasing morbidity rates. This study examines a range of machine learning algorithms such as Naïve Bayes, Decision Tree, Logistic Regression, Random Forest, Neural Network, Support Vector Machine, LogitBoost, and Voting classifier to develop accurate predictive models for diabetes. The data used in this research is drawn from a comprehensive dataset available on mendeley.com, sourced from the laboratory of Medical City Hospital in Iraq. The focus of the study is on feature selection and evaluation metrics to effectively gauge model performance. Eight classification techniques are employed and compared, including Decision Trees (DT), Random Forests (RF), and LogitBoost. The study's findings highlight DT and RF as the top-performing algorithms, demonstrating comparable predictive abilities, with LogitBoost also showing promising results. Conversely, Support Vector Machine (SVM) shows reduced performance due to its sensitivity to outliers. These insights enable healthcare practitioners to adopt appropriate machine learning methods to improve diabetes prediction, thus enabling timely interventions and enhancing patient outcomes.

Keywords: Diabetes, Machine Learning Techniques, Prediction, Comparative Analysis

1. Introduction

Diabetes mellitus is a chronic disease characterized by hyperglycemia. It may cause many complications. According to the growing morbidity in recent years, in 2040, the world's diabetic patients will reach 642 million, which means that one of the ten adults in the future is suffering from diabetes. There is no doubt that this alarming figure needs great attention. With the rapid development of machine learning, machine learning has been applied to many aspects of medical health [1].

Diagnostic and prognostic activities are performed using predictive models in a range of medical disciplines. These models are based on "experience," which is data gathered from real-life situations. The information can be preprocessed and expressed as a collection of rules, as in knowledge-based expert systems, or used as training data for statistical and machine learning models. The purpose of this project is to build a model that will predict whether the patient has diabetes or not depending on some basic features. the dataset is normally collected from lab results. In this study, supervised learning is used to classify Diabetes at an early stage. The findings from this study will enable a better understanding of the classification and regression methods. The findings should also help lay the improvement of the health and awareness.

Diabetes is a group of infections that cause high blood sugar levels. Diabetes is a condition in which the level of glucose in the blood rises. The glucose in our blood comes from the foods we eat on a regular basis. Glucose lingers in our blood if we don't have enough insulin. In the human body, glucose levels can occasionally be higher than usual, but not high enough to be classified as diabetes. However, high glucose levels in human blood can create serious issues. High blood glucose levels can harm the eyes, kidneys, nerves, and cause heart disease, stroke and oral health. Diabetes

***Corresponding author: Hoda A. Abdelhafez (hodaabdelhafez@gmail.com)**
DOI: https://doi.org/10.47738/jads.v5i2.219

is also known as gestational diabetes by pregnant women. this research project is presenting the how the Diabetes is affecting the people and its complication [2].

The aim of this project is to build a model that will predict whether the patient has diabetes or not depending on some basic features. The research objectives focus on how could machine learning techniques help in detecting Diabetes, and how can early diagnoses help the patients.

Diabetes mellitus, also known as diabetes, refers to a chronic and metabolic health condition that influences the body's ability to convert food into energy resulting in high blood sugar [3]. When a patient has diabetes, their body does not generate enough insulin hormones to transport the sugar from the blood system into the cells, or the insulin generated is not properly used. Insulin regulates the sugar levels in the blood [2]. Therefore, the cells stop taking the sugar from the bloodstream leading to sugar saturation in the blood, which may cause serious health challenges such as kidney disease, heart disease, or vision impairment. The symptoms of diabetes were first 1552 B.C by an Egyptian physician Hesy-Ra who documented strange symptoms such as frequent urination and emaciation in patients [4]. Elevated blood sugar levels are a frequent consequence of uncontrolled diabetes, progressively causing significant damage to various bodily systems, particularly the blood vessels and nerves [5].

Globally, there has been a substantial rise in the prevalence of diabetes. According to the World Health Organization (WHO), it was estimated that in 2014, over 422 million individuals were affected by various types of diabetes, a stark increase from 108 million reported in 1980, with most cases concentrated in middle-income and low-income countries [6]. With records as of 2021 estimated to be more than 537 million adults between 20 and 80 years, while projections for the future are about 643 million by 2030 [7]. Saudi Arabia ranked the second in the Middle East, an estimated 4.27 million people live with diabetes, while almost 2 million people are believed to be living with the disease but not been properly diagnosed. There are several types of diabetes including type 1 diabetes, type 2 diabetes, prediabetes, and gestational diabetes [8]. Diabetes is diagnosed after blood screening, which allows medical personnel to detect the amount of sugar in blood.

There are three major types of diabetes, Type 1, Type 2, and gestational diabetes. Type 1 has no major cause and can affect anyone, and previously assumed to affect children only. In Type 1 diabetes, inadequate insulin production necessitates patients to administer insulin into their bodies on a daily basis. Symptoms may start taking place suddenly in a human being and they include constant hunger, fatigue and unexplained weight loss, thirst and constant urination [9]. On the other hand, Type 2 diabetes occurs when the body is unable to effectively use all the insulin that has been produced by the pancreas. It mainly occurs when there is limited body exercise and increased body weight. Since it is very hard to detect the onset of Type 2 disease, it is discovered in patients after a long period and complications associated with it affecting the person. Gestational diabetes, as the name suggests, affects when pregnant, and recover once they deliver. Complications associated with gestational pregnancy include problems while pregnant and while delivering, and the probability of the children developing Type 2 diabetes [7]. Diagnosis takes place through the prenatal screening.

Diabetes causes several complications to the human body such as hyperglycemia toxicity where the bloodstream is saturated with sugar, heart disease, dental decay and gum disease, kidney failure, and other infections that may lead to death. While diabetes has no cure, early diagnosis is important as the patient opportunity to prevent or delay the progression of diabetes. Machine learning (ML) and artificial intelligence (AI) play a significant role in the management of diabetes by enabling the patients to make effective decisions on diet and the level of physical activity required. Early diagnosis is important since anyone can develop diabetes, and with the symptoms being hard to detect, it affirms the need for regular check-up. It involves taking a small sample of blood and testing it to know the level of glucose in the blood. Since Type 1 diabetes cannot be prevented, measures to prevent it can be undertaken, which will prevent the possibility of developing Type 2 [6]. For instance, maintaining a healthy weight based on the body mass index as well as taking a balanced diet and regularly exercising may have a positive impact on the body. The probability of reducing gestational diabetes can also take place by taking healthy meals and maintain moderate weight before pregnancy. Personal responsibility and lifestyle changes will have a direct effect on the ability of a patient to live with diabetes, as long as they are able to follow the medical guidelines.

 ISSN 2723-6471

Complications associated with diabetes stem from the amount of blood sugar in the blood. High sugar levels have a possibility of damaging internal and external organs in a human body, justifying the need for prevention of the negative side effects of diabetes. Chronic complications are the negative side effects that develop over time and include eyesight issues, foot problems, and cardiovascular related problems like heart attack [10]. People with diabetes may have poor eyesight or serious foot problems that result in amputation. Patients may have sexual problems, with men failing to get aroused and women losing sexual sensation and getting regular urinary tract infection. Acute complications often result in chronic complications and depend on sugar level if it is high or low. The complications arise from the unstable high sugar levels in the blood over time, which damages the blood vessels.

2. Related Work

There are some of the previous research studies that illustrated the predictions and classifications models for the diabetics and the complications that accrue mostly with the diabetic patients.

Tan et al. [11] proposed genetic algorithm-stacking ensemble learning model for predicting accurately diabetes risk. The dataset gathered from Qingdao CDC, which contained 8787 desensitization data. To examine the correlation of the attributes in the dataset, a feature correlation heatmap was used which demonstrated the importance of extracting a representative attribute subset. The genetic algorithm (GA) based on decision tree was used for feature selection to improve the accuracy of the model. The authors applied six machine learning algorithms: genetic algorithms. decision tree, support vector machine, K-nearest neighbor, logistic regression Conventional neural network and Naive Bayes. The authors added GA-stacking to each algorithm and compare the results. The comparison includes these six algorithms before and after adding GA through measuring accuracy, precision, F1-score, sensitivity, specificity, and average prediction time. The results demonstrated that using stacking and two primary learners CNN and SVM providing great generalization capabilities of the model. Despite the effectiveness of GA-stacking, the model got limitations in unbalanced data and datasets with small attribute sets.

Pima Indian diabetes dataset of female was used by Kaur and Kumari [12]. To detect risk factors of diabetes, five classification models were applied using R tool. Dataset was about female patients, which collected national institute of diabetes and digestive and kidney diseases. It contained 768 instances with binary classes and eight risk factors. The authors applied outliers, feature selection and predicting missing values using k-nearest neighbor imputation. Boruta Wrapper algorithm was used for feature selection which yielded four important attributes. The implemented algorithms were k-nearest neighbor, neural network, linear kernel and radial basis function, support vector machine, and multifactor dimensionality reduction. The results demonstrated that linear kernel support vector machine and k-nearest neighbor were 0.90 and 0.92 respectively. Thus, these two models were best methods for predicting diabetic patients. Krishnamoorthi et al. [13] used same dataset to introduce framework for predicting diabetes disease. The developed framework was based on machine learning methods. In this dataset, the inconsistent data was removed as well as handling missing values. The authors applied four classification techniques SVM, LR, RF, and KNN. Hyper-parameter tuning was implemented. To select the best hyper-parameter grid search algorithm was used. The results showed that both BMI and glucose had strongly correlation with diabetes. In logistics regression, the percentage of the ROC value was 86%, which was better result compared with the other methods.

Another experiment used Pima Indian diabetes dataset done by Saxena et. al. [14]. This study applied ensemble, stacked ensemble and classical machine learning models. The authors applied feature selection method, which was categorized into three distinct types. Firstly, the filter method employed tests like T-test and chi-square to evaluate the significance of data based on its inherent properties, disregarding dependencies with other features. Secondly, the wrapper method utilized supervised learning algorithms for feature selection. Thirdly, the embedded method employed to search for optimal features. The feature selection showed that four important features which were age, glucose, diabetes ped function and BMI. The authors compared nine algorithms including decision tree, logistic regression, support vector machine, Adaboost classifier, K-nearest neighbor, Linear discriminant analysis, random forest classifier, Gradient boosting classifier and extra tree classifier. A voting ensemble method was applied with the tuned nine machine learning models. The stacked ensemble as a super-learner was applied with two layers: individual models with tuned parameters and single GradientBoost model. The comparison showed that stacked ensemble model provided best accuracy.

Zou et. al. [1] used in their study a dataset collected from the hospital in Luzhou, China. This dataset has 14 attributes and contained 69082 healthy people data and 151598 diabetic data. The study also used Pima Indians diabetics dataset which contained 786 diabetics with 8 attributes and reduced to 392 after removing the missing data. They applied three classification models Random frost, decision tree, and neural network to compare between Luzhou and Pima Indians datasets. Principal component analysis (PCA) and minimum redundancy maximum relevance (mRMR) was used in this study as feature selection methods to reduce the dimensionality. The random forests provided better result compared with the other classifiers. In the Luzhou dataset was 0.8084, and the Pima Indians was 0.7721. Fasting blood glucose in Luzhou dataset and blood glucose tolerance in Pima Indians dataset were used as indicators for prediction based on the three classifiers. The result demonstrated that the fasting glucose feature was significant for predict but depending on this feature do not provide accurate prediction. Moreover, the results showed that using mRMR provided better results compared with PCA, in addition to using all features had better performance in Luzhou dataset.

Dagliati et al. [15] conducted project using machine learning as a branch Artificial Intelligence to predict diabetic patients and its Complications. This project involved 943 patients of type 2 diabetes mellitus, collected from ICSM hospital in Italy. The dataset contained missing values like lipid-related data, therefore the authors applied random forest approach (missForest) and two statistical methods to handle missing data. The missForest had outperformed imputation compared with mean and median statistical methods. After Pre-processing the data, the RF, LR, SVMs and NB algorithms were used on both none-balanced dataset and new balanced dataset using oversampling. The models built based on three microvascular and three temporal thresholds. The LR classifier provided better results with accuracy 0.838, thus it was suitable for predicting the diabetes.

On the other hand, Shin et al. [16] proposed a study based on machine learning to predict diabetes ranging from 2 to 9 years using dataset from Korea. This dataset collected from tertiary hospital in Seoul, and it classified into 1518 diabetic patients and 36,861 nondiabetic patients. The median and modes were used to handle missing values. The authors applied machine learning algorithms via threshold including decision tree, logistic regression, random forest, Cox regression, eXtreme gradient boosting, XGBoost survival embedding and Cox regression. The XGBoost survival embedding algorithm provided better performance model for predicting diabetics and the authors intended to use it in real clinical cases.

A study done by Rajput and Khedgikar [17] used the dataset that was used in this research study. The authors used five machine learning algorithms to predict diabetic patients. The dataset was from was the laboratory of Medical City Hospital in Iraqi. It included 844 diabetic patients, 103 non-diabetic patients and 53 (pre-diabetic patients). The authors applied correlation analysis and feature selection using ANOVA to find features that had impact on determining diabetes. An ANOVA F-test was conducted since the dataset comprises two categorical variables: the diabetes class and gender. The age, HbA1c, VLDL, and BMI Attributes exhibited p-values lower than the alpha value, prompting rejection of the null hypothesis. These four attributes had a substantial impact on determining the diabetes class, thus, they were utilized for model training. The implemented machine learning algorithms were naive bayes, decision trees, multinomial logistic regression, stochastic gradient boosting and Random Forest. The results showed that decision tree and stochastic gradient boosting provided better results. Moreover, they found that the risk for diabetes was high because of the increasing of BMI and age.

3. Research Methodology

This section discusses the appropriate machine learning algorithms for the diabetes dataset to build accurate model for prediction. It reviews feature selection and the evaluation metrics that were applied to the dataset.

3.1. Machine Learning Algorithms for Classification

3.1.1. Naïve Bayes

It is a probabilistic machine learning algorithm utilized for data classification. It employs Bayes' theorem to determine the likelihood of an instance belonging to different classes. The algorithm assumes that the features used for classification are independent, simplifying the calculation of the probability that an instance belongs to a particular class. To classify data, we train the algorithm using a labeled dataset, enabling it to learn the probabilities of each

 ISSN 2723-6471

feature given each class. When presented with a new instance, the algorithm calculates the probability of it belonging to each class based on these learned probabilities [18], [13].

3.1.2. Decision Tree

decision tree is a type of machine learning algorithm that divides data into smaller groups based on the values of input features. They are commonly used for tasks involving classification and regression. The algorithm learns the structure of the tree and criteria for dividing the data during training and can subsequently predict the target variable for new instances by navigating through the tree using their input features [19], [12].

3.1.3. Logistic Regression (RL)

It is a machine learning algorithm used to predict the probability of an instance belonging to a particular class based on its input features. By fitting a logistic function to the training data, the algorithm converts the input features into a probability score ranging from 0 to 1. To use logistic regression for classification, we train the model using labeled data and employ maximum likelihood estimation to learn the parameters of the logistic function. When given a new instance, the algorithm calculates the probability of it belonging to the positive class using the learned logistic function and selects the class with the highest probability [14].

3.1.4. Random Forest

Random Forest is used to enhance the accuracy and reliability of predictions by combining multiple decision trees. Each tree is constructed using a random subset of features and training instances. The final prediction is made by consolidating the predictions from all the trees [15].

3.1.5. Neural Network

Neural Network is a type of machine learning algorithm that consist of interconnected neurons arranged in layers. These networks process input data and make predictions about the output. During training, the network adjusts its weights and biases to minimize a loss function, which measures the difference between its predicted and true labels. Once trained, the network can be used to predict the target variable for new instances [19], [1].

3.1.6. Support Vector Machine (SVM)

It is another popular supervised learning algorithm used for classification and regression tasks. SVMs aim to find the best hyperplane in the feature space by maximizing the margin between the closest points from each class. This is achieved by solving a quadratic optimization problem that involves minimizing a cost function while considering certain constraints [16].

3.1.7. LogitBoost

It is a type of ensemble learning methods that combines boosting and logistic regression techniques to create accurate predictive models with improved performance compared to standalone logistic regression models. The purpose of using EL methods is to obtain a more accurate classification of training data and better generalization on unseen data [20], Boosting is a machine learning technique that sequentially trains weak classifiers and combines their predictions to create a strong classifier. Logistic regression, on the other hand, is a statistical model used to predict binary outcomes. The final prediction of LogitBoost is obtained by combining all weak classifiers' predictions using weighted majority voting. The weights assigned to each weak classifier depend on its performance during training. LogitBoost has several advantages over traditional logistic regression models. It can handle complex interactions between features and automatically select relevant features for classification. It also reduces bias and variance by iteratively adjusting weights and focusing on misclassified samples [21].

3.1.8. Voting Classifier

It is another type of ensemble learning that is used to classify data into different categories based on voting or consensus. These algorithms are commonly used in various applications such as sentiment analysis, spam filtering, and recommendation systems. Vote classification algorithms provide effective solutions for categorizing data into different classes based on voting or consensus among neighboring instances. These algorithms have proven to be versatile and widely applicable in various domains where accurate classification is required [21]. Hard and soft voting are two

methods of voting. Hard voting implements one vote for each stand-alone classifier and then the class label that is selected represents the majority, that has more than half votes. Soft voting uses the average class label probabilities as a voting score, and the final class label has average probability from each classifier or the highest voting score [22].

3.2. Feature Selection

Feature selection, as a component of dimensional reduction, aims to eliminate redundant features and identify an optimal subset from the dataset to build highly accurate models [23]. In this research, two methods were employed: CfsSubsetEval and WrapperSubsetEval. The CfsSubsetEval, also known as Correlation-based feature selection (CFS), selects features highly correlated with the class but uncorrelated with each other [24]. In the CFS method, Genetic Search is one of the search techniques utilized, implementing a search based on genetic algorithms [23]. WrapperSubsetEval, on the other hand, employs an induction algorithm as an evaluation function for identifying a good feature subset, with accuracy estimation techniques measuring the accuracy of induced classifiers [25], [26]. This method employs a wrapper-based selection approach using a genetic algorithm as a search technique to optimize the number of combined attributes that best describe the dataset [27].

3.3. Evaluation Metrics

Evaluation metrics are employed to measure the accuracy of classification models, determining the extent to which the model effectively predicts the correct outcome. Accuracy measurement is calculated as the ratio of correctly classified instances to the total number of instances [28], [29]. Confusion matrix is a table that summarizes the performance of a classification model by showing the counts of true positive, true negative, false positive, and false negative predictions as shown in table 1. It helps in calculating various evaluation metrics such as Precision, Recall, and F1-score [30].

Table 1. Confusion matrix

	Positive	**Negative**
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

Precision measures how many predicted positive instances are actually positive. It calculates the ratio of true positives to the sum of true positives and false positives. Precision focuses on minimizing false positives.

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \quad (1)$$

Recall, also known as sensitivity or true positive rate, measures how many actual positive instances are correctly predicted as positive. It calculates the ratio of true positives to the sum of true positives and false negatives. Recall focuses on minimizing false negatives.

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \quad (2)$$

F1 measure or F1-score is a metric that combines precision and recall into a single score. It provides a balance between these two metrics and is particularly useful when dealing with imbalanced datasets. F1 measure balances precision and recall, precision minimizes false positives, and recall minimizes false negatives. These metrics collectively help in assessing and improving machine learning models' effectiveness in various applications [31].

$$\text{F1} - \text{score} = \frac{2 * (\text{Recall} * \text{Precision})}{(\text{Recall} + \text{Precision})} \quad (3)$$

Accuracy measurement is a metric used to evaluate the performance of a classification model. It measures the proportion of correctly classified instances out of the total number of instances. Accuracy is calculated by dividing the number of correct predictions by the total number of predictions made [29]. The accuracy of the model is determined by the formula below.

 ISSN 2723-6471

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \quad (4)$$

Specificity is a statistical measure that quantifies the ability of a classification model to correctly identify negative instances. In other words, specificity measures the proportion of actual negative instances that are correctly identified as negative by the model [28]. A high specificity value indicates that the model has a low rate of false positives, meaning it accurately identifies negative instances. On the other hand, a low specificity value suggests that the model incorrectly classifies some negative instances as positive. Specificity is calculated as the ratio of true negative (TN) instances to the sum of true negative and false positive (FP) instances:

$$\text{Specificity} = \text{TN} / (\text{TN} + \text{FP}) \quad (5)$$

It is important to note that specificity is complementary to sensitivity (also known as recall or true positive rate). While specificity focuses on correctly identifying negatives, sensitivity measures the ability of a model to correctly identify positives [30], [31].

The Matthews correlation coefficient (MCC) is a measure that used to evaluate the quality of for binary and multiclass classification. The MCC is widely used in machine learning and bioinformatics to assess classification models, particularly when dealing with imbalanced datasets or when there is a significant difference in class sizes [29]. The MCC ranges from -1 to +1, where +1 indicates a perfect prediction, 0 indicates a random prediction, and -1 indicates a completely incorrect prediction [30]. The equation for calculating MCC is as follows:

$$\text{MCC} = \frac{(\text{TP} * \text{TN}) - (\text{FP} * \text{FN})}{\text{Sqrt}((\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN}))} \quad (6)$$

4. Dataset and Analysis

4.1. Diabetic Dataset

The diabetic dataset used in this research study was available in mendeley.com. It was gathered from the laboratory of Medical City Hospital in Iraqi (the Specializes Center for Endocrinology and Diabetes-Al-Kindy Teaching Hospital). [32]. The data were extracted from patients' files and stored in database. The data included medical information and laboratory analysis. The datasets included age, gender and 9 laboratory attributes as shown in table 2. The dataset of 1000 patients are classified into diabetics (y), non-diabetics (N), and pre-diabetics (P).

Table 2. Dataset attributes

No.	Attribute	Description
1	Gender	Male /female
2	AGE	20 - 79
3	Urea	Blood Urea level
4	Cr	Creatinine ratio
5	HbA1c	Hemoglobin A1C
6	Chol	Cholesterol
7	TG	Triglycerides
8	HDL	High-density lipoprotein
9	LDL	Low-density lipoprotein
10	VLDL	Very-low-density lipoprotein
11	BMI	Body Mass Index
12	CLASS	Diabetic, Non-diabetic, or Predict-diabetic

4.2. Data Pre-Processing

The quality of the medical data is affected by missing values, outliers, and other factors. Therefore, data pre-processing is used to focus on these factors before applying the machine learning algorithms. Two steps were used to preprocess the data:

4.2.1. Missing Data

Refers to the data values that are not stored for a variable or some variables in the dataset [33]. Missing medical data causes significant problem in the data analysis process [34]. We examined the dataset to determine the presence of missing data and considered the possibility of using imputation techniques via the ReplaceMissingValues filter in Weka to substitute any missing values with statistically derived estimates, such as the mean or median. However, upon inspection, we discovered that the dataset did not contain any missing data.

4.2.2. Outlier

Represents anomalous patterns in data that are not in the normal behavior rang. Detecting these anomalous patterns or error outliers helps in handling them to provide accurate prediction model especially for machine learning algorithms. If it is an error outlier then simply remove this entry from the dataset [35], [36]. To find the outliers in the dataset, one of the used techniques is the interquartile range (IQR). In this technique, the calculated Inter Quartile Range represents the variation in the data. It is a difference between first quartile and third quartile; IQR = Q3 − Q1. The IQR interquartile range is used if the data point that fall outside the range of [Q1 +1.5*IQR] and [Q3 + 1.5*IQR] are considered outliers. Thus, any value that lies outside the range is noted as an outlier where IQR represents 75th percentile – 25th percentile [36]. To identify outliers within the dataset, the Interquartile Range filter in Weka was applied, facilitating the detection and removal of such instances. This filter operated in an unsupervised manner. It calculated the interquartile range (IQR) for each numeric attribute in the dataset and filters out instances that fall outside a specified range, thus aiding in the identification and removal of outliers. This process contributed to enhancing the dataset's integrity, mitigating the potential impact of outliers on subsequent analysis or modeling outcomes.

5. Implementation and Result

5.1. Implementation

Three main phases were mentioned previously in the research methodology section. The first phase was collecting and understanding the diabetes dataset. The second phase was pre-processing the data before applying classification models. The third phase was applying feature selection and then applying 8 machine learning algorithms for prediction as shown in figure1. These algorithms were Naive Bayes, Decision Trees, Logistic Regression, Support Vector Machines, Neural Networks, Random Forest, logitboost and voting. The dataset was split into 70 % training set and 30 % for testing set. The prediction classification involved whether the patient is Diabetes, pre-diabetes or not diabetes.

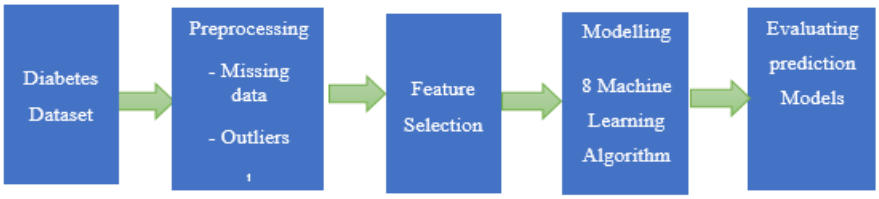

Figure 1. Implementation phases

Following data pre-processing, two feature selection techniques were employed: CfsSubsetEval and WrapperSubsetEval. CfsSubsetEval focuses on selecting features with high predictive power and low redundancy, while WrapperSubsetEval directly evaluates the performance of the model with different feature subsets to find the optimal combination of features that maximizes predictive accuracy. The first utilized method was CfsSubsetEval with GeneticSearch, a genetic algorithm for attribute selection. In CfsSubsetEval with a genetic search, the algorithm commenced with a dataset comprising randomly generated feature subsets. Each subset underwent evaluation based on the CfsSubsetEval criteria, which entailed assessing the correlation between each feature and the class, as well as

the correlation between features themselves. Subsequently, the subsets were ranked according to their fitness, favoring those demonstrating higher correlation with the class and lower redundancy. The genetic algorithm then applied genetic operators to generate new feature subsets. These operations encompassed mutation, which randomly altered some features within a subset, and crossover, combining features from two parent subsets to generate new offspring subsets. The newly created subsets were evaluated using CfsSubsetEval, and this iterative process persisted until a predetermined stopping criterion was satisfied. The outcome revealed seven attributes highly correlated with the class: gender, age, HbA1c, Chol, TG, VLDL, and BMI. The second method was WrapperSubsetEval with genetic algorithm. It initiated with a dataset containing randomly generated feature subsets. Each subset underwent evaluation using an induction algorithm (classifier) to assess its performance. Subsequently, genetic operators such as mutation and crossover were employed to generate new feature subsets, and this process iterated until a stopping criterion was fulfilled WrapperSubsetEval with GeneticSearch, identified four highly correlated attributes with the class: gender, HbA1c, TG, and BMI. Both CfsSubsetEval and WrapperSubsetEval contributed to optimizing the model by selecting the most informative subset of features.

Subsequently, the eight prediction models were applied to all attributes, as well as to the subsets of seven and four correlated attributes, aiming to determine the optimal prediction results. In voting classifier as an ensemble method combined the prediction outputs of six classifiers: Naive Bayes, Decision Trees, Logistic Regression, Support Vector Machines, Neural Networks, Random Forest. These specified algorithms were selected for diabetes prediction due to the following reasons:

1) Naïve Bayes is adept at forecasting diabetes based on a range of risk factors like age, gender, BMI, and blood glucose levels
2) Decision Tree proves valuable in diabetes prediction, as it can handle interactions among various factors such as age, weight, and blood sugar levels in intricate ways
3) Logistic Regression offers insights into the most influential features for predicting diabetes likelihood
4) Random Forest demonstrates high effectiveness, especially when dealing with numerous features and their potential interactions, making it suitable for predicting a multifaceted condition like diabetes
5) Neural Network proves valuable in predicting diabetes, especially when the relationships among variables are complex
6) Support Vector Machine (SVM) exhibits strength in diabetes prediction, particularly when specific combinations of factors distinctly indicate the presence or absence of the condition.
7) LogitBoost is beneficial for enhancing prediction accuracy
8) Voting Classifier, in diabetes prediction, leverages the strengths of different algorithms, as various algorithms may excel in capturing different aspects of the data, leading to more robust and accurate predictions when their predictions are combined through voting.

5.2. Results and Discussion

The following tables represent the results of applying the eight prediction models. All features were used for predicting diabetes without removing outlies and the results are shown in table 3 and figure 2. The results indicate that the DT and RF have the best and demonstrated similar results among the eight classifiers as well as logitboost, while the SVM model is considered the worst because of its sensitivity to the outliers.

Table 3. Predicting diabetes all features and without removing outliers.

ML	Accuracy	Sensitivity	Specificity	F1	MCC
DT	99.33%	0.993	0.983	0.993	0.976
SVM	87.33%	0.873	0.416	0.840	0.496
RF	99.33%	0.993	0.967	0.993	0.976

ISSN 2723-6471

NN	96.33%	0.963	0.849	0.962	0.874
LR	91.33%	0.913	0.763	0.907	0.718
NB	94.00%	0.940	0.945	0.943	0.813
logitboost	99.00%	0.990	0.966	0.990	0.964
Voting	97.67%	0.977	0.900	0.976	0.915

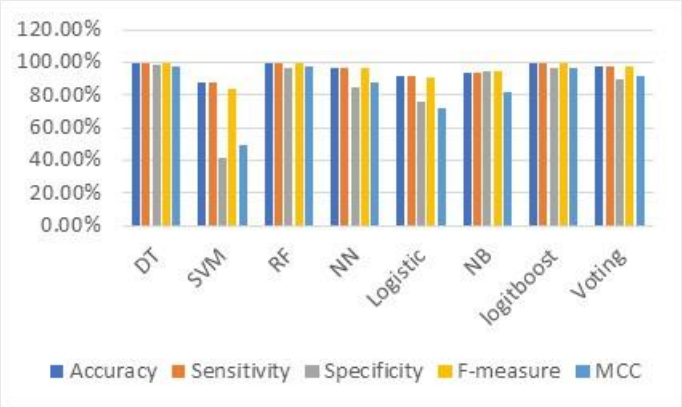

Figure 2. The results of predicting diabetes without removing outliers

According to the table 4 and figure 3, show the results of the prediction based on all feature and with removing outliers. The results demonstrated that DT has a better performance compared to the other models, followed by RF and logitboost that have same accuracy 97.54. On the other hand, the accuracy of SVM was enhanced because of removing the outliers.

Table 4. Predicting diabetes all features with removing outliers

ML	Accuracy	Sensitivity	Specificity	F1	MCC
DT	97.89%	0.979	0.960	0.979	0.923
SVM	90.18%	0.902	0.519	0.883	0.596
RF	97.54%	0.975	0.941	0.975	0.908
NN	95.09%	0.951	0.882	0.950	0.842
LR	91.93%	0.919	0.766	0.910	0.712
NB	92.00%	0.916	0.917	0.921	0.741
logitboost	97.54%	0.975	0.960	0.976	0.911
Voting	96.84%	0.968	0.902	0.968	0.880

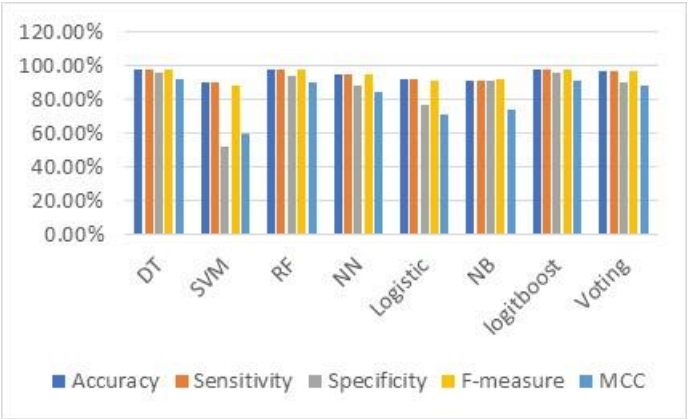

Figure 3. The results of predicting diabetes with removing outliers

ISSN 2723-6471

The results of using four features: gender, HbA1c, TG and BMI are showing in table 5 and figure 4. These results indicate that DT and Vote have approximately the same results and superior performance among other prediction models.

Table 5. Predicting diabetes based on 4 features.

ML	Accuracy	Sensitivity	Specificity	F1	MCC
DT	98.33%	0.983	0.917	0.983	0.940
SVM	95.00%	0.950	0.832	0.948	0.824
RF	97.67%	0.977	0.883	0.976	0.916
NN	97.00%	0.970	0.915	0.970	0.893
LR	90.00%	0.900	0.631	0.876	0.612
NB	97.00%	0.967	0.947	0.967	0.894
logitboost	97.67%	0.977	0.916	0.976	0.916
Voting	98.00%	0.980	0.976	0.981	0.918

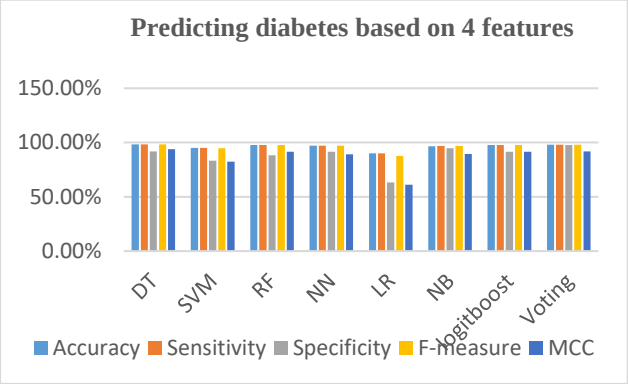

Figure 4. The results of predicting diabetes with Four selected features

Table 6 and figure 5 demonstrate the results of applying 7 feature which are gender, age, HbA1c, Chol, TG, VLDL, and BMI. We found the RF provides better performance than the other models followed by DT that provides better diabetes prediction.

Table 6. Predicting diabetes based on 7 features

ML	Accuracy	Sensitivity	Specificity	F1	MCC
DT	99.33%	0.993	0.983	0.993	0.976
SVM	91.00%	0.910	0.681	0.902	0.687
RF	99.67%	0.997	0.983	0.997	0.988
NN	97.00%	0.970	0.866	0.969	0.899
LR	92.33%	0.923	0.747	0.917	0.719
NB	95.00%	0.953	0.930	0.955	0.851
logitboost	98.67%	0.987	0.950	0.986	0.952
Voting	98.33%	0.983	0.933	0.983	0.940

According to table 7 and table 8, both methods of using 7 selected features and all features (without removing outliers) have a better accuracy and ACC for six prediction models. The other two models NB and SVM provide best accuracy and MCC in 4 selected features compared with the other methods. Figure 6 and figure 7 show the accuracy and MCC for eight prediction models respectively.

ISSN 2723-6471

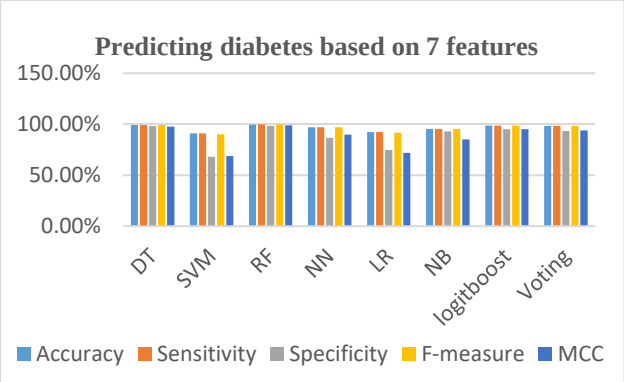

Figure 5. The results of predicting diabetes with 7 selected features

Table 7. TNE accuracy of 8 machine learning algorithms

ML	all features without removing outliers	all features with removing outliers	Four features	Seven features
DT	99.33%	97.89%	98.33%	99.33%
SVM	87.33%	90.18%	95.00%	91.00%
RF	99.33%	97.54%	97.67%	99.67%
NN	96.33%	95.09%	97.00%	97.00%
LR	91.33%	91.93%	90.00%	92.33%
NB	94.00%	92.00%	97.00%	95.00%
logitboost	99.00%	97.54%	97.67%	98.67%
Voting	97.67%	96.84%	95.00%	98.33%

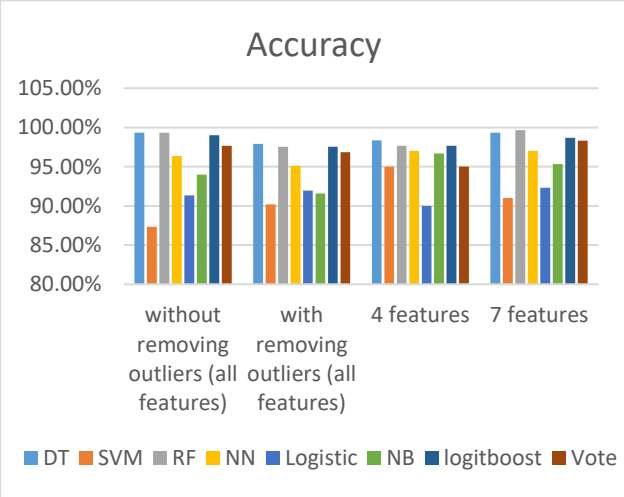

Figure 6. The results of applying of 8 Machine learning algorithms using accuracy measure

Table 8. The MCC of 8 machine learning algorithms

ML	without removing outliers (all features)	with removing outliers (all features)	Four features	Seven features
DT	0.976	0.923	0.940	0.976
SVM	0.496	0.596	0.824	0.687
RF	0.976	0.908	0.916	0.988
NN	0.874	0.842	0.893	0.899

ISSN 2723-6471

Logistic	0.718	0.712	0.612	0.719
NB	0.813	0.741	0.894	0.851
logitboost	0.964	0.911	0.916	0.952
Voting	0.915	0.880	0.824	0.940

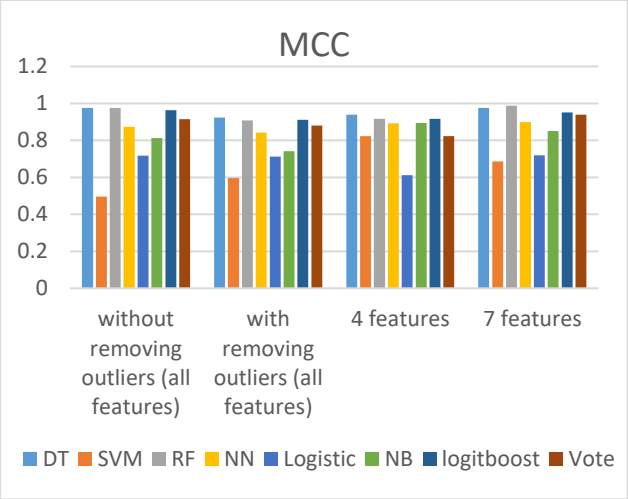

Figure 7. The results of applying of 8 Machine learning algorithms using MCC measure

Table 9 demonstrates that both DT and RF exhibit the highest levels of accuracy, F1 scores, and MCC, along with significant sensitivity and specificity. Collectively, these findings imply that DT and RF stand out as the most effective algorithms for diabetes prediction when utilizing all features and seven selected features from this dataset. Conversely, NB and SVM yield the best accuracy, sensitivity, specificity, F1 scores, and MCC when utilizing the four selected features, in comparison to other methods. Thus, the selection of features can significantly impact the performance of machine learning models.

Table 9. Summarize the results of 8 machine learning algorithms

	Metric	DT	SVM	RF	NN	LR	NB	Logit-Boost	Voting
All Features (without outliers)	Acc	0.993	0.873	0.993	0.963	0.913	0.940	0.990	0.976
	Sen.	0.993	0.873	0.993	0.963	0.913	0.940	0.990	0.977
	Spec.	0.983	0.416	0.967	0.849	0.763	0.945	0.966	0.900
	F1	0.993	0.840	0.993	0.962	0.907	0.943	0.990	0.976
	MCC	0.976	0.496	0.976	0.874	0.718	0.813	0.964	0.915
All Features (with outliers)	Acc	0.978	0.901	0.975	0.950	0.919	0.92	0.975	0.968
	Sen.	0.979	0.902	0.975	0.951	0.919	0.916	0.975	0.968
	Spec.	0.96	0.519	0.941	0.882	0.766	0.917	0.960	0.902
	F1	0.979	0.883	0.975	0.950	0.910	0.921	0.976	0.968
	MCC	0.923	0.596	0.908	0.842	0.712	0.741	0.911	0.880
4 Features	Acc	0.983	0.950	0.976	0.970	0.900	0.970	0.976	0.980
	Sen.	0.983	0.950	0.977	0.970	0.900	0.967	0.977	0.980
	Spec.	0.917	0.832	0.883	0.915	0.631	0.947	0.916	0.976
	F1	0.983	0.948	0.976	0.970	0.876	0.967	0.976	0.981
	MCC	0.940	0.824	0.916	0.893	0.612	0.894	0.916	0.918
7 Features	Acc	0.993	0.910	0.996	0.970	0.923	0.95	0.986	0.983
	Sen.	0.993	0.910	0.997	0.970	0.923	0.953	0.987	0.983
	Spec.	0.983	0.681	0.983	0.866	0.747	0.93	0.950	0.933
	F1	0.993	0.902	0.997	0.969	0.917	0.955	0.986	0.983
	MCC	0.976	0.687	0.988	0.899	0.933	0.851	0.952	0.940

 ISSN 2723-6471

Table 10 demonstrates the comparison with the reference [17]. This reference selected 5 features that were gender, age, VLDL, HbA1c and BMI. The authors applied Decision tree, Random forest, logistic regression, and Naive Bayes. The results indicated that the prediction models in our study outperformed the models examined in the study done by Rajput and Khedgikar.

Table 10. Comparison with another reference using selected features

ML	Proposed Model Four features	Proposed Model Seven features	Reference Five features
DT	98.33%	99.33%	95.07%
SVM	95.00%	91.00%	
RF	97.67%	99.67%	90.64%
NN	97.00%	97.00%	
Logistic	90.00%	92.33%	86.7%
NB	97.00%	95.00%	93.1%
logitboost	97.67%	98.67%	
Voting	95.00%	98.33%	

6. Conclusion

In conclusion, the alarming rise in diabetes prevalence underscores the critical need for accurate predictive models to aid in early diagnosis and intervention. Leveraging machine learning techniques offers promising avenues for detecting diabetes based on essential patient features. This study employed comprehensive data pre-processing and feature selection methodologies to enhance the quality of the analysis.

The evaluation of eight prediction models revealed that Decision Trees, Random Forests, and LogitBoost consistently demonstrated superior performance in predicting diabetes, while Support Vector Machine exhibited sensitivity to outliers, impacting its accuracy. Notably, the removal of outliers notably improved the accuracy of SVM, emphasizing the importance of robust data pre-processing techniques.

Further analysis considering different feature subsets highlighted the efficacy of DT and RF, particularly when utilizing seven selected features. These models consistently outperformed others in terms of accuracy and Matthew's correlation coefficient. Naive Bayes and SVM exhibited competitive accuracy when utilizing four selected features.

Overall, the findings suggest that Random Forests and Decision Trees offer robust predictive capabilities for diabetes detection, especially when considering a comprehensive set of features. These models hold promise for facilitating early diagnoses, enabling timely interventions, and ultimately improving patient outcomes in the face of the growing diabetes epidemic. Future research could explore additional feature engineering techniques and ensemble methods to further enhance predictive performance and clinical utility. Moreover, the future work could be broadened to encompass specific directions for further investigation, such as exploring deep learning models, integrating additional patient data, or implementing real-time prediction systems in clinical settings.

7. Declarations

7.1. Author Contributions

Conceptualization: H.A.A. and A.A.A.; Methodology: H.A.A. and A.A.A.; Software: H.A.A.; Validation: H.A.A. and A.A.A.; Formal Analysis: H.A.A. and A.A.A.; Investigation: H.A.A.; Resources: A.A.A.; Data Curation: A.A.A.; Writing Original Draft Preparation: H.A.A. and A.A.A.; Writing Review and Editing: A.A.A. and H.A.A.; Visualization: H.A.A.; All authors have read and agreed to the published version of the manuscript.

7.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

7.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

7.4. Institutional Review Board Statement

Not applicable.

7.5. Informed Consent Statement

Not applicable.

7.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju and H. Tang "Predicting Diabetes Mellitus With Machine Learning Techniques", *Front. Genet*, vol. 9:515, pp. 1-10, doi: 10.3389/fgene.2018.00515, 2018.

[2] World Health Organization, 10 November 2021 [Online]. Avalable: https://www.who.int/news-room/fact-sheets/detail/diabetes.

[3] J. B Cole and J. C Florez "Genetics of diabetes mellitus and diabetes complications", *Nature reviews nephrology*, vol. 16, no. 7, pp. 377-390, 2020.

[4] N. Khan, N. Alam, K. Eqbal and G. Nahid, "Historical account of diabetes-An Overview", *The Pharma Innovation Journal*, vol. 9, no, 9, pp. 26-30. 2020.

[5] N. Barakat, A. Bradley and M. Barakat, "Intelligible support vector machines for diagnosis of diabetes mellitus", *IEEE transactions on information technology in biomedicine: a publication of the IEEE Engineering in Medicine and Biology Society*, vol. 14, no.4, pp. 1114-1120, 2010.

[6] World Health Organization. (n.d.). Diabetes. World Health Organization, 5 April 2023, [Online]. https://www.who.int/news-room/fact-sheets/detail/diabetes

[7] J. Bell. Prevalence of diabetes in Saudi Arabia to almost double by 2045: Report. Al Arabiya English. Retrieved March 26, 2022, from https://english.alarabiya.net/News/gulf/2021/11/15/Prevalen e-of-diabetes-in-Saudi-Arabia-to-almost-double-by-2045-Report., November 15, 2021

[8] O. Jacqmain, IDF webinar: Different types of diabetes–June 21, 2019-Webinar: Living with, 2019.

[9] J. L. Harding, M. E. Pavkov, D. J. Magliano, J. E. Shaw and E. W. Gregg, "Global trends in diabetes complications: a review of current evidence", *Diabetologia*, vol. 62 no. 1, pp. 3-16, 2019.

[10] E. Brutsaert E., Complications of Diabetes Mellitus, MSD Manual for the Consumer, https://www.msdmanuals.com/home/hormonal-and-metabolic-disorders/diabetes-mellitus-dm-and-disorders-of-blood-sugar-metabolism/complications-of-diabetes-mellitus, 2023

[11] T. Yaqi, C. He, Z. Jianjun, T. Ruichun and L. Peishun, "Early Risk Prediction of Diabetes Based on GA-Stacking", *Applied Sciences*, vol. 12, no 2, p. 632, 2022.

[12] H. Kaur and V. Kumari, "Predictive modelling and analytics for diabetes using a machine learning approach", *Applied Computing and Informatics*, vol. 18, no. 1/2, 2022, PP. 90-100, 2022.

[13] R. Krishnamoorthi, S. Joshi, H. Almarzouki , P. Shukla, A. Rizwan, C. Kalpana, and B. Tiwari, "A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques", *Journal of Healthcare Engineering*, vol. 2022, no. 1, pp. 1-10, 2022

[14] S. Surabhi, M. Debashish, P. Subhransu and S. Kumar, "Machine learning algorithms for diabetes detection: a comparative evaluation of performance of algorithms," *Evolutionary Intelligence*, vol. 16, no 2, pp. 587-603, 2021.

[15] A. Dagliati, S. Marini, L. Sacchi, G. Cogni, M. Teliti, M. Tibollo, P. De Cata, L. Chiovato, and R. Bellazzi, "Machine Learning Methods to Predict Diabetes Complications", *Journal of Diabetes Science and Technology*, vol. 12, no. 2, pp. 295–

302, 2018.

[16] J. Shin et al., "Improving Machine Learning Diabetes Prediction Models for the Utmost Clinical Effectiveness," *J. Personalized Medicine*, vol. 12, no. 11, pp. 2-10, 2022.

[17] R. Minakshi and K. Sushant, "Diabetes prediction and analysis using medical attributes: A Machine learning approach," *Journal of Xi'an University of Architecture and Technology*, vol. XIV, no. 1, pp. 98-103, 2022.

[18] Y. Tan, H. Chen, J. Zhang, R. Tang, and P. Liu, "Early Risk Prediction of Diabetes Based on GA-Stacking," *Applied Sciences*, vol. 12, no. 2, art. no. 632, pp. 1-14, 2022.

[19] M. R. Rajput and S. S. Khedgikar, "Diabetes prediction and analysis using medical attributes: A Machine learning approach," *Journal of Xi'an University of Architecture and Technology*, vol. XIV, no. 1, pp. 98-103, 2022.

[20] P. Yang, Y. H. Yang, B. B. Zhou, and A. Y. Zomaya, "A Review of Ensemble Methods in Bioinformatics," *Curr. Bioinform.*, vol. 5, no. 4, pp. 296–308, 2016.

[21] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research," *Computational and Structural Biotechnology Journal*, vol. 15, no. 1, pp. 104–116, 2017.

[22] A. Yuniarti and M. A. Fauzi, "Ensemble method for Indonesian twitter hate speech detection," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 11, no. 1, pp. 294–299, 2018.

[23] A. Onik, N. Haq, and L. Alam, "An Analytical Comparison on Filter Feature Extraction Method in Data Mining using J48 Classifier," *International Journal of Computer Applications*, vol. 124, no. 1, pp. 1-8, 2015.

[24] M. A. Hall, "Correlation-based Feature Subset Selection for Machine Learning," University of Waikato, 1999.

[25] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1–2, pp. 273-324, Dec. 1997.

[26] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings IJCAI-95, Montreal, Que.*, vol. 1, no. 1, pp. 1137-1143, 1995.

[27] M. Sainin, R. Alfred, A. F. Ahmad, and M. Lammasha, "An Evaluation of Feature Selection Methods on Multi-Class Imbalance and High Dimensionality Shape-Based Leaf Image Features," *Journal of Telecommunication, Electronic and Computer Engineering*, vol. 9, no. 1-2, pp. 57-61, 2017.

[28] N. Japkowicz and M. Shah, Evaluating Learning Algorithms: A Classification Perspective, Cambridge University Press, 2011.

[29] D. M. Powers, Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation," *Journal of Machine Learning Technologies*, vol. 2, no 1, pp. 37-63, 2011.

[30] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing and Management*, vol. 45, no. 4, pp. 427-437, 2009.

[31] M. L. McHugh, "Interrater reliability: The kappa statistic," *Biochemia Medica*, vol. 22, no 3, pp. 276-282, 2012.

[32] A. Rashid, "Diabetes Dataset," Mendeley Data, v1, 2020. [Online]. Available: doi: 10.17632/wj9rwkp9c2.1. [Accessed: March 20, 2024].

[33] H. Kang, "The prevention and handling of the missing data," *Korean J. Anesthesiol*., vol. 64, no. 5, pp. 402–406, May 2013.

[34] P. Royston, "Multiple Imputation of Missing Values,"*Stata J. Promot. Commun. Stat. Stata*, vol. 4, no. 3, pp. 227–241, Aug. 2004.

[35] H. P. Vinutha, B. Poornima, B. M. Sagar "Detection of Outliers Using Interquartile Range Technique from Intrusion Dataset". In: S. Satapathy, J. Tavares, V. Bhateja, J. Mohanty (eds) Information and Decision Sciences. Advances in Intelligent Systems and Computing, vol 701. Springer, Singapore. https://doi.org/10.1007/978-981-10-7563-6_53, 2018

[36] C. Dash, A. Behera, S. Dehuri, A. Ghosh "An outliers detection and elimination framework in classification task of data mining", *Decision Analytics Journal,* Vol. 6, no 2, PP. 1-8, ISSN 2772-6622, https://doi.org/10.1016/j.dajour.2023.100164, 2023.