Al-Nisour Journal for Medical Sciences

Manuscript 1160

Predicting the Risk of Diabetes Using Machine Learning Algorithms

Noor Ismail Ibrahim

Jamal Kamil Alrudaini

Hytham Falih Hassan

Sahar Faeq Jaafar

AL-NISOUR JOURNAL FOR MEDICAL SCIENCE 2026;8:24–38

Predicting the Risk of Diabetes Using Machine Learning Algorithms

Noor Ismail Ibrahim [a], Jamal Kamil Alrudaini [b,*], Hytham Falih Hassan [b], Sahar Faeq Jaafar [c]

[a] Computer Engineering Techniques, Al-Nisour University, Baghdad, Iraq
[b] Cybersecurity Engineering Techniques, Al-Nisour University, Baghdad, Iraq
[c] Anesthesia Techniques Department, Al-Nisour University, Baghdad, Iraq

Abstract

Diabetes is a chronic disorder that many people suffer from. This disease develops when the blood glucose level is high. Serious side effects, such as harm to the heart, kidneys, eyes, and other organs may result from diabetes neglected treatment. Diabetes has numerous causes, including aging, obesity, inactivity, genetics, poor food and lifestyle choices. Early identification of this illness helps lessen its negative consequences. There are many traditional methods for predicting this disease, but they are expensive. Early prediction of diabetes benefits all those at risk by providing early treatment. With the advancement of healthcare technology, machine learning algorithms can analyze large amounts of data, which can help the medical sector make more accurate and timely decisions. In this paper, artificial intelligence algorithms were used to help medical professionals predict this disease, as these technologies can greatly help the medical sector by predicting the possibility of diabetes with the utmost accuracy, thus saving time for both doctors and patients. The main goal of this study focused is to employ machine learning algorithms to analyze medical data and select the best algorithm for predicting the disease by comparing the evaluation metrics of these algorithms. The Indian diabetes dataset PIMA obtained from the Irvine Machine Learning (ML) Repository in the University of California, was used. This research applied four algorithms including the decision tree algorithm (C4.5), K-Nearest Neighbor algorithm, random forest algorithm, and support vector machine algorithm was used to predict diabetes. The experiment results illustrated that the random forest algorithm did the best overall than other models in predicting disease.

Keywords: K-Nearest Neighbors (K-NN), Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), and Machine Learning

1. Introduction

Diabetes is an ongoing disorder that takes place in the body of humans when the body is unable to consume insulin effectively or when the pancreas is not producing sufficient of it. Insulin as a hormone is in charge of regulating blood sugar levels. All types of diabetes increase the risk of early death and can lead to complications in a number of body areas. A leg amputation, kidney failure, stroke, nerve damage, heart attack, and blindness are among the potential side effects. The risk of fetal death and other problems is also increased by poorly managed diabetes during pregnancy.

The World Health Organization reported that there were more than 400 million cases of diabetics in 2014, which is 108 million up from cases reported in 1980. It is reported to be the reason of death for around 1.5 million in 2012, while high blood glucose levels were thought to be the cause of an additional 2 million deaths. The occurrence of diabetes is likely to be raised from 500m in 2021 to 1.31 billion in 2050, making it a significant global public health concern. Type 2 diabetes reported to represent 96% of the diabetes cases reported in 2021 (Khanam & Foo, 2021).

The healthcare industry collects a massive quantity of information incorporating medical records of

Received 21 December 2025; accepted 12 March 2026.
Available online 20 May 2026

* Corresponding author.
E-mail addresses: noor.i.eng@nuc.edu.iq (N. I. Ibrahim), jamal.k.eng@nuc.edu.iq (J. K. Alrudaini), haytham.f.cyb@nuc.edu.iq (H. F. Hassan), sahar.f.anesth@nuc.edu.iq (S. F. Jaafar).

https://doi.org/10.70492/2664-0554.1160

patient and test results. illness prediction is examined using the doctor's experience and knowledge for early illness diagnosis, but this can be imprecise and prone to errors; therefore, manual decisions can be upsetting and influence decision-making, depriving patients of the right therapy (Sheng et al., 2024).

AI technologies can contribute to healthy living by monitoring diabetic patient and providing personalized advice and plans based on individuals' health needs and lifestyle, especially in pre-diabetes or early diabetes stages. AI revolutionize diabetes care by predicting risks and personalizing treatment plans to improve patient outcomes (Olisah et al., 2022).

AI analyzes large data sets to predict risks, accurately interpret test results, personalize treatment plans, and enhance outcomes for patients and care. There is significant promise to lessen the burden of diabetes management by implementing AI into the healthcare system and moving toward a more proactive and individualized approach (Khalifa & Albadawy, 2024).

This research is distinguished from other reported studies by using the SMOTE technique to address issues of imbalanced classes in the dataset, which led to improving the accuracy of the models' performance and increasing their comprehensiveness. According to the investigation, the Random Forest algorithm's accuracy was 99.35%, which reflects the quality of the models used and the efficiency of their data processing.

2. Related works

Artificial intelligence methods have been applied in several fields, especially in the healthcare field. We quote some works that use artificial intelligence algorithms to solve medical problems.

Researcher (Sisodia & Sisodia, 2018) conducted a study to create an approach that has the ability to forecast the likelihood of patient facing diabetes. The Pima Indian Diabetes Dataset (PIDD), a dataset collected by researchers at the University of California, and published publicly on Irvine's Machine Learning Repository, was analyzed using three machine learning classification techniques for this purpose: decision trees, support vector machines, and Naive Bayes. This study's primary goal is to use the PIDD medical database and the Waikato Environment for Knowledge Analysis (WEKA) technology for forecasting patients with diabetes.

The three algorithms' performances are assessed using several measures, including recall, accuracy, precision, and F1-measure. Both correctly and erroneously classified occurrences are used to gauge accurateness. The results showed that the Naive Bayes performs better comparing to the other presented methods, with the an 76.30% of accuracy and high Receiver Operating Characteristic (ROC) curves.

Another investigation of ML techniques in the area of diabetes estimation was carried out by Khanam & Foo (2021) features like glucose, BMI, insulin, pregnancy, and age were incorporated and one output feature (score) were used in the PIMA data set then the data was cleaned and prepared using the WEKA tool and three features were eliminated using the feature reduction approach. To forecast the development of diabetes and assess efficiency on several measures, the researchers employed seven ML algorithms: DT, KNN, RF, NB, ANN, LR, and SVM. According to the results, the neural network technique outperformed other machine learning classifiers and had the best accuracy of all the models that were put forth. With an accuracy rate of almost 86%, the neural network with two hidden layers confirmed to be the effective and promising for the analysis of diabetes.

In (Khanam & Foo, 2021) Machine learning algorithms including K-Nearest Neighbor, Decision Trees, Random Forest, AdaBoost, Naive Bayes, and XGBoost and Multilayer Perceptron (MLP) were used to build a classification model for diabetes early detection using the Indian Pima diabetes dataset. Since there should be little correlation between the base classifiers, the results indicated that the combination of two boosted classifiers named AB and another one called XB is the optimal amalgamation for diabetes detection. When the suggested preprocessing and correlation-based feature selection are used, the optimal combination (AB+XB) can produce the highest diabetes prediction accuracy. Furthermore, the findings demonstrate that the developed model can assist medical professionals in detecting diabetic patients more accurately.

In the research work (Yahyaoui et al., 2019), a decision support system was applied to predict diabetes using ML techniques and deep learning (DL) algorithms, where traditional ML algorithms such as SVM and RF were applied, Conversely, for DL, a fully convolutional neural network (CNN) was used to detect and predict patients with diabetes. The findings from the experiments indicate that RF was more successful in predicting diabetes than SVM and deep learning techniques, indicating the usefulness of the random forest algorithm in assisting medical professionals in making disease predictions.

Six distinct AI techniques leveraged in the discussion of predictive analytics in healthcare in Sarwar et al. (2018). LR, DT, RF, KNN, SVM, and NB are some of these algorithms. The PIMA dataset, which is a dataset created by researchers at the University of California and made publicly

available by Irvine Machine Learning Repository, is used to make diabetes predictions. The objective of the proposed system is to use AI approaches to assist medical professionals in the early detection of diabetes. According to the testing results, KNN and SVM algorithms predict diabetes with the best accuracy when assessed with the four other methods.

In (Nahzat & Yağanoğlu, 2021) an algorithm is applied to classify the possibility of having diabetes, to achieve the goal the authors used five ML classification methods namely artificial neural networks, decision tree, k-nearest neighbor algorithm, random forest, and support vector machine to predict diabetes. Preparing and implementing diabetes detection using different methods of machine learning and analyzing their outputs to detect the effective and reliable classifier with the best accuracy is the primary goal of this study. The results indicate that, out of the various algorithms used, the random forest algorithm produces the best results.

In (Permana et al., 2021) diabetes detection system is designed with the goal of which is to identify the influential variable among several variables that cause diabetes. The machine learning notion served as the grounds for the construction of the technique that was presented, by applying the decision tree algorithm (C4.5) to predict diabetes to help people in charge analyzing and detecting the disease early. The results showed that heavy drinking plays a role in diabetes and with higher accuracy using the decision tree.

In (Islam et al., 2020) the role of artificial intelligence methods using random forest algorithm is discussed, provides a foundation for determining a patient either has diabetes or not, depending on diabetes risk factors. This study's goal is to accurately identify the diabetes type using random forest algorithm, using a dataset collected from Khulna Diabetes Center, Bangladesh. The dataset contains two types of typical and atypical symptoms. The results showed that artificial intelligence methods have the ability to predict the disease with higher accuracy using random forest, which can identify potential patients in the future for prevention of the disease.

In (Mujumdar & Vaidehi, 2019) a classification model for diabetes prediction is applied. The authors employed different ML methods, including classification, clustering, and regression. A statistical computing tool for diagnosing diabetes was R-Studio software. The University of California Irvine (UCI) repository provided the Indian PIMA database. Improving the predictive model's accuracy is the primary goal of this research. The results show that SVM and logistic regression algorithms give optimal results among all the algorithms used.

In (Santhanam & Padmavathi, 2015), K-Means is used to remove any noise in the data. After that the genetic algorithms was applied to detect the optimal set of attributes by incorporating SVM to train on the database for diabetes prediction by using diabetes dataset obtained from UCI repository, the results achieved using the presented approach give better accuracy related to modified data preparation method based on K-Means clustering with SVM classifier.

The summary of the state of the art is as shown in the Table 1 below.

3. Diabetes disease

This type of disorder is considered as an ongoing disorder that arises once the human body is unable to effectively utilize insulin. produced by body or when the organ is unable to produce the amount required. The hormone that is in charge of blood sugar levels and controlling it in the body is insulin. High blood sugar, or hyperglycemia, is a frequent consequence of untreated diabetes. It has the potential to seriously harm numerous bodily systems over time, particularly those involving blood vessels and neurons. Diabetes cannot be treated, on the other hand this disease can be managed with a healthy diet, frequent exercise, medication, and monitoring for complications (Panda et al., 2022).

3.1. Types of diabetes disease

A variety of diabetes types that can differ from each other in terms of the mechanism of occurrence and the nature of the symptoms in addition to the treatment (Sonar & JayaMalini, 2019):

3.1.1. Type 1 diabetes

The body in such a kind of diabetes try to resist directly and kills the pancreas beta cells that generate the insulin, which is brought on by an autoimmune reaction. Because insulin is necessary for cells to consume glucose from the blood, the body will be unable to create it, and this will result in elevated blood glucose levels. The term juvenile diabetes or childhood diabetes refers to the fact that type 1 diabetes typically manifests in children, particularly those between the ages of 4 and 14. Type 1 diabetes indicators can range in severity from moderate to severe, and they usually appear rapidly. This variety is characterized by symptoms such as weight loss, poor learning and growth in children, seizures, and loss of consciousness brought on by a sharp increase in blood sugar.

Table 1. The state of the art is as follows.

Previous studies	Advantages	Disadvantages
Prediction of Diabetes using Classification Algorithms	The study helps identification of diabetes, enabling patients to receive prompt treatment and reduce the risk of complications.	Despite the good performance, the accuracy of 76.30% is not ideal, meaning there is a significant amount of error in the prediction.
A comparison of machine learning algorithms for diabetes prediction (Sisodia & Sisodia, 2018)	The study highlights the use of advanced tools like WEKA to analyze important data, which helps in finding out relevant things easily and effectively.	Some AI algorithms lack explanation, such as neural networks, as they operate in a way that is difficult for humans to interpret or understand.
Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers (Khanam & Foo, 2021)	Ensembling different models (such as XGBoost and AdaBoost) helps reduce overall errors, as the strengths of different models are combined to achieve the best performance. This increases accuracy, sensitivity, and specificity, which helps in early detection of diabetes.	Despite the use of cross-validation and grid search, the model may face the problem of overfitting (especially if multiple models are used) if the data is limited or not diverse enough. This means that the model may learn the fine details in the training data but fail to take a broad view of them to new coming data.
A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques (Yahyaoui et al., 2019)	The study showed that Random Forest (RF) was the most effective in identifying diabetes with 83.67% of an accuracy of, which is an outstanding result that can help doctors diagnose the disease more accurately.	Although there were unbalanced classes in the dataset (268 diabetic patients vs. 500 non-diabetic), no processing techniques were reported to balance the data or improve model performance in unbalanced cases.
Prediction of Diabetes Using Machine Learning Algorithms in Healthcare (Sarwar et al., 2018)	A comprehensive comparative study of six commonly used machine learning algorithms is presented, helping in choosing the appropriate method for healthcare.	In this study, the highest accuracy was (77%), which is not considered sufficient to be relied upon in multiple applications, as medical decision-making requires high accuracy.
Diabetes Prediction Using Machine Learning Classification Algorithms (Nahzat & Yağanoğlu, 2021)	Through comparing different techniques such as KNN, RF, DT, SVM, and ANN, the study provides insights into the most efficient and accurate algorithms, helping in choosing the best technique for similar applications.	Artificial network models are considered as deep learning and need big data to train the model well.
Classification of diabetes disease using decision tree algorithm (C4.5) (Permana et al., 2021)	The model used achieved a rate of 90.38%, which proved its effectiveness in diagnosing diabetes.	Although the C4.5 algorithm showed good results, this study did not compare it with other algorithms such as random forest or logistic regression that may be more efficient.
Typical and Non-Typical Diabetes Disease Prediction using Random Forest Algorithm (Islam et al., 2020)	The notable results of the study is that it provides an effective way to detect diabetes in the advanced stage, which reduces its long-term effects such as effects on the eyes, heart and all nerves.	While the model may be accurate in this study, it may face challenges when applied in the real world, especially if the input data is inconsistent or if there are large changes in patient control or treatment.
Diabetes Prediction using Machine Learning Techniques (Mujumdar & Vaidehi, 2019)	This study provides an effective way to predict early stage diabetes, allowing doctors and healthcare professionals to prevent the disease early. This could lead to reduced complications of the disease such as heart problems and blood pressure.	Achieving a 77% rate means a 23% chance of error, in a healthcare context, causing unnecessary anxiety for the patient and prompting unnecessary treatments or tests.
Application of K-Means and Genetic Algorithms for Dimension Reduction by Integrating SVM for Diabetes Diagnosis (Santhanam & Padmavathi, 2015)	The study found that the proposed model, which combines K-Means clustering techniques, genetic algorithms (GA) for dimensionality reduction and feature selection, and SVM for classifying the diabetes dataset, achieved an accuracy of 98.79%. This performance represents a significant improvement over traditional methods including combining K-Means and SVM, which showed an accuracy of 96.71%.	The study focused on one metric, precision, however, other important metrics for instance precision, recall, and F1-measure were not specified, which may give an incomplete picture of the system performance.

3.1.2. Type 2 diabetes

For type 2, the resistance of insulin, which is the method that controls the production of insulin, becomes unable to use it efficiently which causes unbalance and lead to type 2 diabetes. In addition to genetic and environmental variables, a number of lifestyle habits, including being overweight and a lack of activity can raise the chance of acquiring this kind of diabetes.

Although type 2 diabetes continues to occur over time and people with it may not experience any symptoms during the initial phases of the disease, its symptoms are similar to those of type 1 diabetes.

Since this type of diabetes typically affects adults, particularly those over 45, commonly identified as adult-onset diabetes. However, due to an inactive lifestyle and rising rates of obesity and gaining weight, this kind has lately started to afflict younger people as well, particularly children and adolescents.

3.2. Gestational diabetes

During pregnancy, this form of diabetes develops when a woman's blood sugar levels start to increase even if they were normal in the days leading up to the pregnancy, as the name implies.

Pregnancy diabetes is influenced by genes and being overweight. If neglected, gestational diabetes might lead to a worse effect on the baby and the mother's healthy life. Due to that, it's critical to regularly examine blood sugar concentrations and administer the right treatment if the mother is diagnosed with the condition.

3.3. Risk factors for diabetes disease

The type of diabetes and history in the family determines the risk factors for the disease. Geographical location and environmental factors can potentially raise the incidence of type 1 diabetes. Autoantibodies, which are immune cells that fight diabetes, are occasionally tested for in family members of persons with type 1 diabetes.

Another factor that may raise the likelihood of type 2 diabetes is belonging to a certain race or ethnicity. Although the exact cause is unknown, some groups are more vulnerable than others, such as Asian Americans, Black people, Hispanics, and American Indians. Additionally, those who are overweight or obese are more likely to have type 2 diabetes, prediabetes, and gestational diabetes (Bouillon et al., 2013).

4. Methodology used

4.1. Algorithms used

4.1.1. Decision tree (DT) algorithm

The DT algorithm is a type of ML algorithm that leverages a chain of decisions determined by values in a set of parameters to analyze data and categorize it into groups (or classifications). These decisions can be seen as tree structure, with the input split into Leaves (sub-branches) and classifications, and decisions being made according to certain criteria. The decision tree algorithm is classified as an algorithm that is supervised, and the decision tree algorithm used widely in tasks such as classification and regression. That are one of the effective techniques frequently employed in several domains, including machine learning, image processing, data mining, and statistics (Mienye & Jere, 2024). A DT approach can be utilized to build a training model for predicting the target's class or value. variables through acquiring decisions from previous data (training data). In contrast with other classification techniques, the decision tree algorithm is incredibly simple. This approach uses a tree model to try to answer the problem, where an attribute is represented by each inner node, a decision rule by each branch, and the result by each terminal node. This hierarchical model mimics human decision-making processes, making it intuitive and easy to interpret (Patel & Prajapati, 2018).

Several kinds of DT algorithms exists including: Binary Iterators 3 (ID3), CART Classification and Regression Tree, ID3 Successor (C4.5), Chi-squared Automatic Interaction Detector (CHAID), Generalized and Unbiased Interaction Detection and Estimation (GUIDE), MARS Multivariate Adaptive Regression Splines, CRUISE Classification Rule with Unbiased Interaction Selection, CTREE Conditional Inference Trees, and Estimation, QUEST Fast, Unbiased and Efficient Statistical Tree (Mienye & Jere, 2024). The DT's top node, referred to as the "Root Node," holds the complete data collection and is separated into two or more homogeneous groups. The sub-node that is further divided into other sub-nodes is known as the "Decision Node". The nodes that do not divide are called the "Leaf Node" and are considered the end of this branch of the tree (Kumar et al., 2012).

4.1.2. C4.5 algorithms

The C4.5 is an enhanced version of the ID3 algorithm presented in 1993 by Quinlan Ross, and this algorithm is one of the types of algorithms that work in the style of a decision tree, as it works to classify cases into different categories. This algorithm falls

under the group of supervised classifiers, which classifies certain cases into a number of categories using the divide-and-conquer method, as it divides the complex problem into simpler problems, then the same function is called automatically for all parts of the problem, and by collecting the solutions of the divided problems, the solution to the complex problem is reached (Priyam et al., 2013; Patidar et al., 2015; Pandey & Jain, 2017).

It is better than ID3 algorithm because it can handle continuous and categorical features. In C4.5, the partitioning is done based on information gain and the feature with the best gain of information by making the decision node and is partitioned further. C4.5 deals with over-fitting by pruning, i.e. it removes branches/sub-parts from the tree that are not of great importance or are redundant. C4.5 algorithm follows post-pruning, i.e. removing branches after the tree is constructed. C4.5 uses information gain ratio which is an indicator of entropy shift; the more information gained, the lower the entropy.

Definition 1 (Information Entropy)**:** The entropy of a training set T is defined as follows when the target attribute takes n distinct values:

$$\text{Entropy}\,(T) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

where the probability is represented by Pi that T belongs to class i.

Definition 2 (Information Gain)**:** The information acquisition of a property A, relative to a set of examples T, is:

$$\text{InfoGain} = \text{Entropy}\,(T) - \text{Entropy}\,(A, T)$$

$$\text{InfoGain} = \text{Entropy}\,(T) - \sum_{i=1}^{n} \frac{|T_i|}{|T|} \text{Entropy}\,(T_i)$$

where Si is the division of S resulting from the value of property A.

Definition 3 (Gain Ratio)**:** The gain ratio "normalizes" information gain as follows (Patidar et al., 2015):

$$\text{GainRatio}\,(A, T) = \frac{\text{InfoGain}\,(A, T)}{\text{SplitEntropy}\,(A, T)}$$

$$\text{GainRatio}\,(A, T) = \frac{\text{InfoGain}\,(A, T)}{-\sum_{i=1}^{n} \frac{|T_i|}{|T|} \log_2 \frac{|T_i|}{|T|}}$$

4.1.3. KNN algorithm

Among the most significant and straightforward algorithms for supervised machine learning is the KNN algorithm. In tasks involving regression and classification, it is also one of the algorithms utilized. Additionally, it can handle high values or abnormal information with great efficiency 22. The principle of operation of this algorithm is derived from calculating the Euclidean distance among points, where the closer two coordinates are to one another, the more probable it is that the two coordinates are part of each other, hence the name of this algorithm. The letter K refers to the number of samples that will classify a point based on the distances between it and its neighbors. The KNN algorithm is used because of its ease of interpretation and short computation time. Manhattan, Minkowski, and Euclidean distances are the three often used formulas for determining the separation between two points (Thant et al., 2020).

The length of a straight line between two places is known as the Euclidean distance. In k-dimensional Euclidean space, the distance between points indicated by Cartesian coordinates is generally:

$$\text{Euclidean Distance} = D\,(X, Y) = \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2}$$

Manhattan Distance: it is the distance between two city blocks (mathematically two vectors) is equal to one standard distance between the vectors.

$$\text{Manhattan Distance} = D\,(X, Y) = \sum_{i=1}^{n} |X_i - Y_i|$$

Minkowski distance: It is a standard vector space that is a generalization of both the Euclidean distance and the Manhattan distance.

$$\text{Minkowski Distance} = D\,(X, Y) = \left(\sum_{i=1}^{n} |X_i - Y_i| \right)^{\frac{1}{p}}$$

However, it may happen that the data we are dealing with does not only contain fixed values but may also contain categorical values. In this case, the method used to calculate the distance will be the Hamming interval. This type of calculation aims to find the number of bit positions in which the two bits differ. The number assigned to K affects the performance of the classification and its ability to deal with different data. If the value of K is small (K=1), the classification will be more affected by noise or outliers in the data, which may lead to inaccurate classification. While a large value of K can reduce the random effect or noise in the data and make the classification more

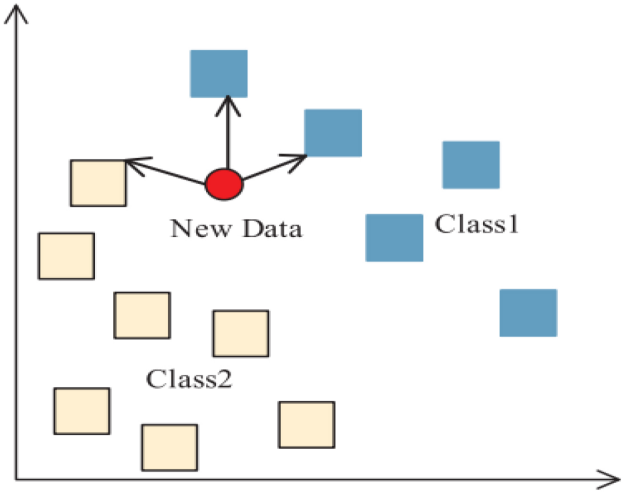

Fig. 1. KNN nearest neighbor algorithm.

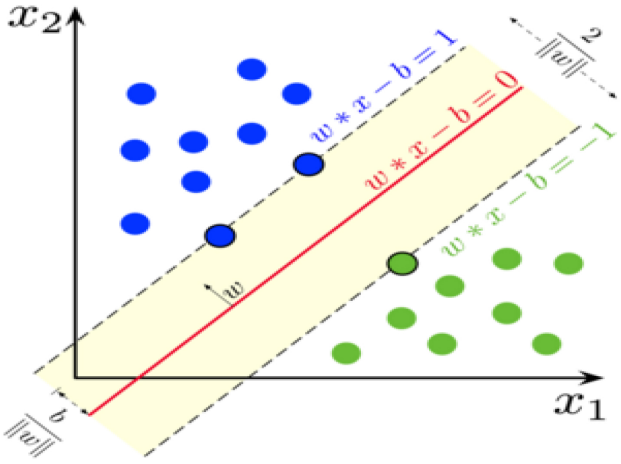

Fig. 2. The hyper-plane that separates the data in the SVM algorithm.

stable. However, increasing the value of K can lead to the loss of some fine details in the data and reduce the ability to recognize fine differences between classes24. Fig. 1 shows the KNN algorithm.

4.1.4. SVM algorithm

This supervised ML technique can be leveraged in situations involving both classification as well as regression. Each sample of data is characterized as an individual feature in a space with n dimensions by the SVM method, where number of features equal to n. The value of the feature is characterized by a distinctive coordinate. Next, we classify by identifying the hyper-plane that best separates the two distinct classes. The SVM algorithm's primary goal is to locate the decision boundary (hyper-plane) in the feature space that most effectively splits the several classes. This hyper-plane is chosen to be responsible for increasing the distance between the adjacent objects of every class. These points closest to the hyper-plane are referred to as support vectors. The classification becomes more accurate the further the points are from the hyper-plane, so these points must be as far from the hyper-plane as possible while remaining on the correct side of the dividing line. These points are very important for the hyper-plane as they determine the best existing hyper-plane. Deleting the support vectors will result in the formation of a new hyper-plane. To obtain the superior hyper-plane We must have the greatest margin. The margin of error is the separation from the hyper-plane and the support vectors. The larger the distance, the better the classification performance (Çil et al., 2020; Yu et al., 2010; Ovirianti et al., 2022).

In pattern recognition, the linear discriminant function in n-dimensional space is:

$$g(x) = w.x + b$$

The hyper-plane equation can be written as:

$$g(x) = (w.x) + b = 0$$

In the case of linear separation, the normalization is applied to the discriminant function g(x), where $g(x)|\geq 1|$ of all training samples are satisfied until they meet away from the sample surface classification. Due to that, the class interval is equivalent to $\frac{2}{||w||}$ and so the class interval becomes equivalent to $||w||$ or $||w||^2$, when doing a classification of the surface of all correctly classified samples it is necessary to satisfy:

$$Y, [(w.x_i) + b] - 1 \geq 0, i = 1, 2, \ldots, n$$

Then, make $||w||^2$ the minimum classification surface, which is the optimum classification surface. Because they support the ideal classification surface, the points on the hyper-plane are referred to as support vectors. Fig. 2 shows the hyper-plane (Ovirianti et al., 2022).

4.1.5. Random forest

This supervised ML approach is applicable to classification as well as regression issues. The bootstrap aggregating technique, which involves selecting at random the training data with replacement to build numerous subgroups of the data, which applied to generate collections of decision trees. Each subgroup is used to train a DT, and the final forecast is constructed by adding up all of the trees' predictions. However, because it can manage enormous volumes

of incomplete and multiple dimensions' data, Random Forest is a powerful technique. It may be applied to multi-class and binary classification applications and performs well with unbalanced datasets (Ibrahim & Jaber, 2022; Soni & Varma, 2020; Alrudaini et al., 2022).

The ensemble learning concept, which is the fusion of multiple classifiers, is the basis of this methodology. to address this complicated issue and advance the accuracy of model, several DTs are put together in the learning method as an ensemble forest to guarantee predictions that are more precise. Each DT is trained using a random sampling of data as well as a random array of features. To create a final forecast, the algorithm then aggregates the predictions from each tree. Because the approach uses numerous decision trees, each of which is trained on a random subset of data, it is known as a "random forest." Making several DTs and integrating their forecasts is how random forest operates. A random subset of features and an arbitrary portion of data are chosen at the start of the method. This subset of data is then used to build a decision tree. Given that each DT is learned using a unique collection of information and attributes, this procedure is carried out several times. Following the creation of each DT, the algorithm aggregates the predictions of each DT to produce a final forecast (Sheng et al., 2024).

4.2. *Evaluation metrics*

Evaluating artificial intelligence algorithms is essential for any project. In this research, several criteria were employed to test suitable models, including precision, accuracy, F1-score, and recall. The significance of these metrics lies in the fact that they offer a consistent method for assessing how well the model predicts or classifies data depending on input (Fida et al., 2011; Abdulsahib et al., 2025; Abbas et al., 2024).

4.2.1. *Confusion matrix*

A table that compares the number of correct and incorrect forecasts that the model for classification produced to the actual outcomes is called the confusion matrix, and it is used to assess the model's performance. The real classes in the matrix are represented by each row, whereas the expected classes are represented by each of the columns. The number of classes determines the size of the matrix. Table 2 shows the confusion matrix:

TP: Expresses the number of actual positive cases that the model correctly predicted as positive (example: TP indicates patient samples correctly classified as abnormal meaning the patients have diabetes disease).

Table 2. Shows the confusion matrix.

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

TN: refers to the count of true negative cases accurately identified by the model.

(example: TN refers to the number of normal people with a normal diabetes condition).

FP: indicates the actual negative cases that the model incorrectly labeled as positive.

(example: FP indicates the number of samples that were incorrectly detected as having diabetes disease).

FN: refers to the count of actual positive instances misclassified as negative by the model. (example: FN refers to people who actually have diabetes disease but the disease is still not detected by the system).

From the obtained confusion matrix, we can determine the levels of Accuracy, Precision, Recall and F-Score as follows (Acharya, 2017):

4.2.2. *Accuracy*

It is a commonly utilized technique for evaluating model effectiveness. It is determined by dividing the overall amount of samples by the proportion of really categorized samples (true positive, true negative).

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

4.2.3. *Precision*

The proportion of actual positive instances that receive a positive classification (i.e. the probability that a patient who receives a positive screening test has the disease).

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

4.2.4. *Recall*

Its definition is the percentage of positive cases that are appropriately labeled as such, the recall ratio is simply a measure of the number of true samples that were predicted from all samples.

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

4.2.5. *F1-score*

F1-scoreis described as a metric that attempts to communicate the harmony between recall and precision. A classifier's performance in terms of precision and recall is shown by its F1 score. Thus, the in order to obtain an improved F1 score, it is beneficial to have a greater precision and recall value. This equation can

Table 3. Explanation of the data set.

Sequence	Attribute	Definition
S. 1	Pregnancies	Number of times pregnant
S. 2	Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
S. 3	Blood Pressure	Diastolic blood pressure (mm Hg)
S. 4	Skin Thickness	Triceps skin fold thickness (mm)
S. 5	Insulin	2-Hour serum insulin (mu U/ml)
S. 6	BMI	Body mass index (weight in kg/(height in m)^2)
S. 7	Diabetes Pedigree Function	Diabetes pedigree function
S. 8	Age	Age (years)
S. 9	Outcome	Class variable (0 or 1)

be used to compute it:

$$\text{F1} - \text{Measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} \times \text{Recall}}$$

4.3. Dataset description

In this research, the Indian Diabetes Dataset (PIMA) was used. This dataset was originally derived from the National Institute of Diabetes and Digestive and Kidney Diseases. This dataset contains 9 attributes and 768 cases, 268 of which have diabetes and 500 have healthy conditions. Table 3 contains a description of the dataset.

4.4. Data preprocessing

This step is a crucial step in ML because it guarantees that the data is cleaned and prepared for the model to learn effectively. Data preparation can increase such models' precision and consistency, leading to better predictions and more accurate decisions. The following preprocessing steps were implemented:

4.4.1. Data cleaning

While there are no null data in the Indian diabetes dataset, a number of anomalies were discovered in several characteristics that may compromise the accuracy of the model. These are some values from the dataset which are very different from the others. This interquartile range approach was employed in this study to exclude outliers. The variation that exists between an organized data set's upper and lower quartile is known as the spread (or dispersion) measure. In a box plot, it is a crucial element.

4.4.2. Data balancing

The dataset used in this research contains 268 diabetic samples and 500 healthy cases. This is considered as an unequal distribution of classes within the dataset and is one of the main reasons for low accuracy. The SMOTE oversampling technique was applied to prevent over fitting and information loss. Using the feature space commonalities among the current marginal cases, synthetic data is created by this approach. It locates every minority case's K nearest neighbors, chooses individual arbitrary. After that, it performs linear interpolations to generate a new minority case in the neighborhood to construct a synthetic case.

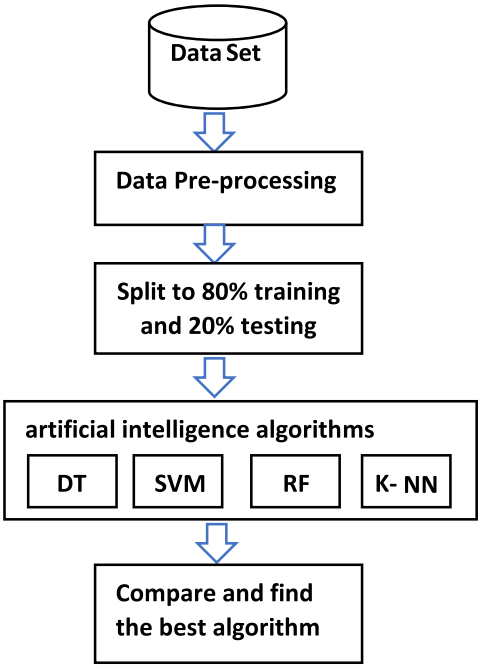

Fig. 3. Shows the system architecture.

5. System architecture

System architecture provides an overview of how the system works, and the operation of this system is described as shown in Fig. 3.

Table 4. Accuracy metrics for the C4.5 decision tree algorithm.

Class	Precision	Recall	F-Score	Accuracy
0	0.81	0.77	0.79	**73.37%**
1	0.62	0.67	0.64	

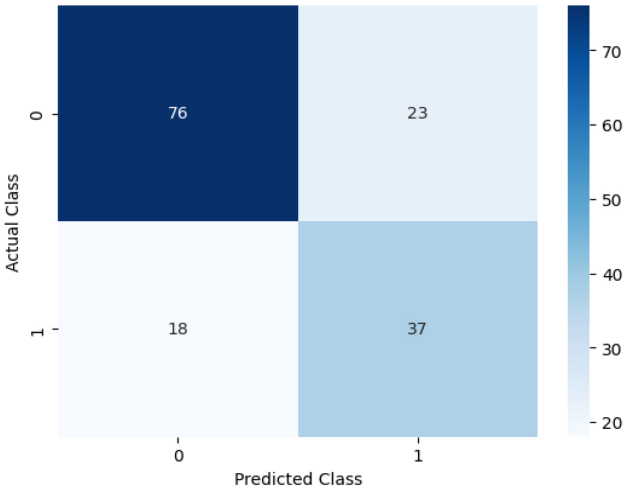

Fig. 4. Confusion matrix of C4.5 algorithm.

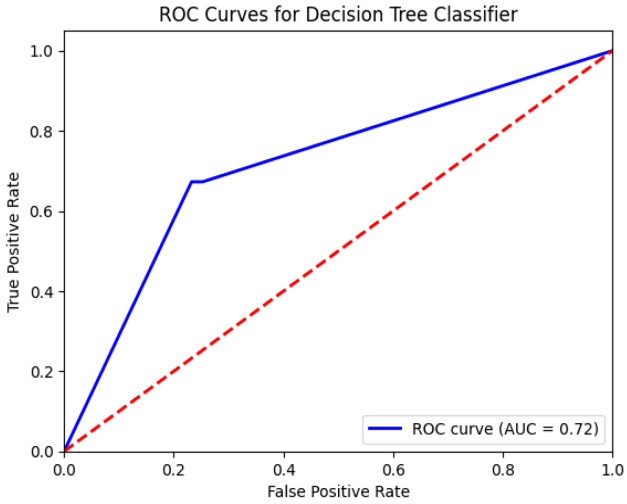

Fig. 5. ROC curve and AUC value for C4.5.

Table 5. Accuracy metrics for the SVM algorithm.

Class	Precision	Recall	F-Score	Accuracy
0	0.65	0.96	0.77	**63.63%**
1	0.43	0.05	0.10	

6. Results and discussion

6.1. Before using SMOTE

6.1.1. Analysis of the results of the C4.5 algorithm

Table 4 shows the accuracy measures for the C4.5 algorithms, where the accuracy of this algorithm reached 73.37%, while the accuracy rate for the positive category for the uninfected case reached 0.81 and the infected case reached 0.62, while the recall rate for the positive category for the uninfected case reached 0.77 and the infected case reached 0.67, while the F-Score value was 0.79 for the uninfected case and 0.64 for the infected case.

Fig. 4 shows the confusion matrix results for the C4.5 algorithms for classifying diabetes into two classes, where the main diagonal indicates the cases that were correctly classified, while the remaining values indicate the cases that were misclassified. The number (23) indicates the positive cases that were misclassified as negative, while the number (18) indicates the negative cases that were misclassified as positive.

Fig. 5 shows the ROC curve for the C4.5 algorithms. This curve is used to measure the performance of classification models and shows the relationship between the true positive rate and the false positive rate. The ROC curve is used to calculate the AUC score. The AUC value ranges from 0 to 1, and the higher the AUC, the better the model performs in distinguishing between positive and negative classes.

From Fig. 5, we notice that the AUC value for this algorithm is 0.72, indicating that the model achieves good performance in distinguishing between the two classes, but it is not ideal.

6.1.2. Analysis of the results of the SVM algorithm

Table 5 shows the accuracy measures for the SVM algorithm, where the accuracy rate for this algorithm was 63.63%, and it shows that the accuracy rate for the positive class for the uninfected case was 0.65 and the infected case was 0.43, while the recall rate for the positive class for the uninfected case was 0.96 and the infected case was 0.05, while the F-Score value was 0.77 for the uninfected case and 0.10 for the infected case.

Fig. 6 shows the confusion matrix results for the SVM algorithm for classifying diabetes into two classes, where the number (4) indicates positive cases that were misclassified as negative, while the number (52) indicates negative cases that were misclassified as positive.

From Fig. 7, we note that the AUC value for the SVM algorithm is 0.69, indicating the model's ability to distinguish between the two classes, but it is not good.

6.1.3. Analysis of the results of the Random Forest algorithm

Table 6 shows the accuracy measures of the Random Forest algorithm, where it is noted that the accuracy of the Random Forest algorithm is 74.02%, and it shows that the accuracy rate for the positive class for the

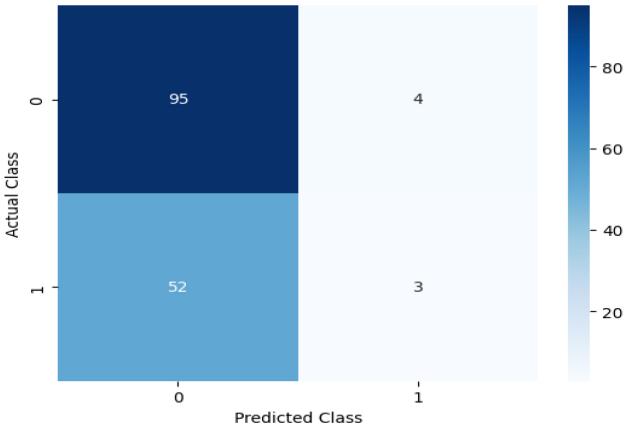

Fig. 6. Confusion matrix of SVM algorithm.

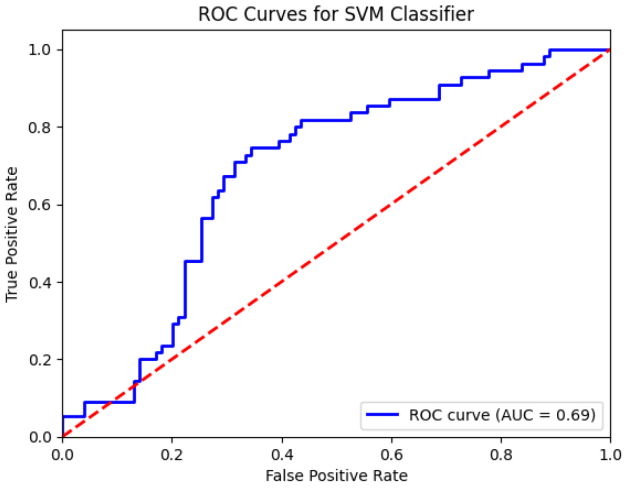

Fig. 7. ROC curve and AUC value for SVM.

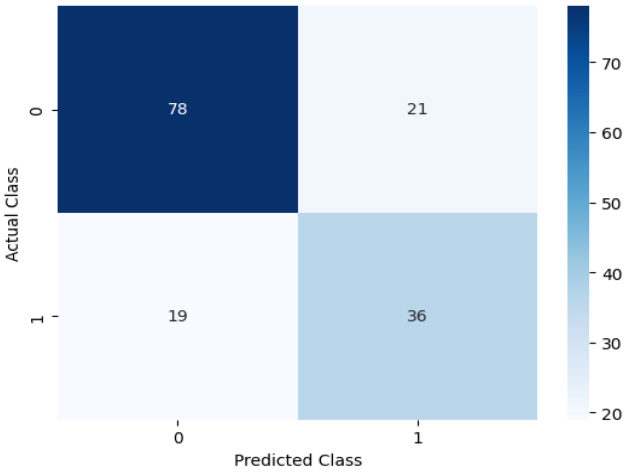

Fig. 8. Confusion matrix of RF algorithm.

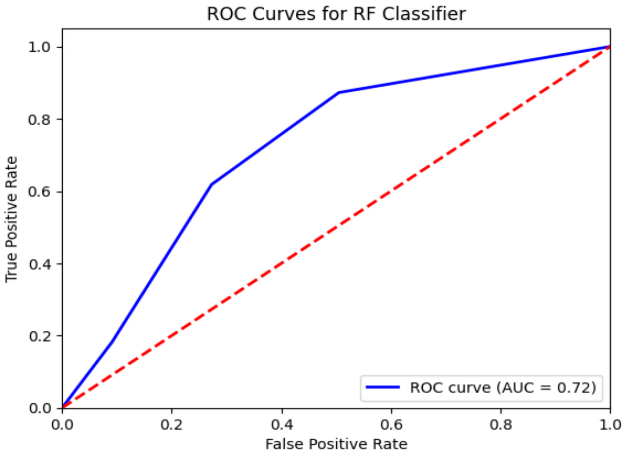

Fig. 9. ROC curve and AUC value for RF.

Table 6. Accuracy metrics for the Random Forest algorithm.

Class	Precision	Recall	F-Score	Accuracy
0	0.80	0.79	0.80	**74.02%**
1	0.63	0.65	0.64	

Table 7. Accuracy metrics for the K-NN algorithm.

Class	Precision	Recall	F-Score	Accuracy
0	0.77	0.73	0.75	**68.83%**
1	0.56	0.62	0.59	

uninfected case is 0.80 and the infected case is 0.63. Regarding the recall rate for the positive class, it was 0.79 for the uninfected case and 0.65 for the infected case, while the F-Score value for the uninfected case was 0.80 and 0.64 for the infected case.

Fig. 8 shows the confusion matrix results for the RF algorithm for classifying diabetes into two classes, where the number (21) indicates positive cases that were misclassified as negative, while the number (19) indicates negative cases that were misclassified as positive.

We note from Fig. 9 that the AUC value of the RF algorithm is 0.72, indicating that the model achieves good performance in distinguishing between the two classes, but it is not ideal.

Analysis of the results of the K-NN algorithm:

Table 7 shows the accuracy measures for the K-NN algorithm, where it is noted that the accuracy of this algorithm is 68.83%, and it shows that the accuracy rate for the positive class for the uninfected case is 0.77 and the infected case is 0.56, while the recall rate for the positive class for the uninfected case is 0.73 and the infected case is 0.62, and with regard to the F-Score value, it was 0.75 for the uninfected case and 0.59 for the infected case.

Fig. 10 shows the confusion matrix results for the K-NN algorithm for classifying diabetes into two

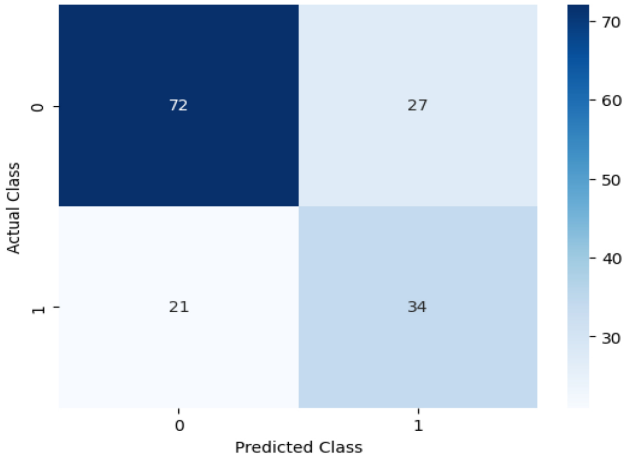

Fig. 10. Confusion matrix of K-NN algorithm.

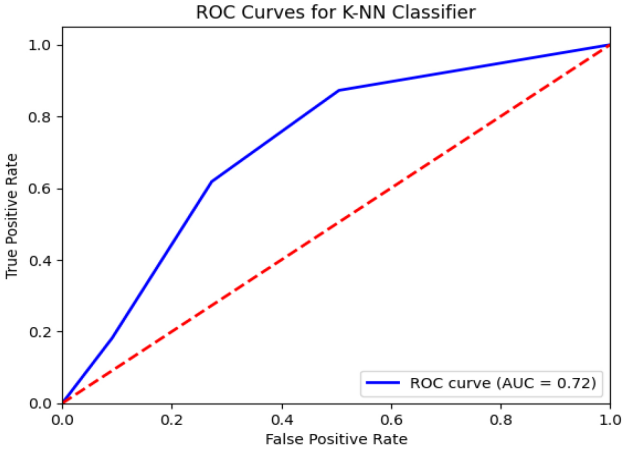

Fig. 11. ROC curve and AUC value for K-NN.

categories, where the number (27) indicates positive cases that were misclassified as negative, while the number (21) indicates negative cases that were misclassified as positive.

Fig. 11 shows that the AUC value of the K-NN algorithm is 0.72, indicating that the model achieves good performance in distinguishing between the two classes, but it is not ideal.

6.2. After using SMOTE

In this study, SMOTE technology was used to obtain the best results. A significant improvement in the results of the algorithms used is observed, demonstrating the effectiveness of this technique in improving the performance of models in imbalanced classification problems.

Table 8 shows the accuracy metrics for the C4.5 algorithms, showing an increase in the accuracy metrics of this algorithm after using SMOTE.

Table 8. Accuracy measures of the C4.5 algorithms after using the SMOTE technique.

Class	Precision	Recall	F-Score	Accuracy
0	0.98	1.00	0.99	**98.70%**
1	1.00	0.96	0.98	

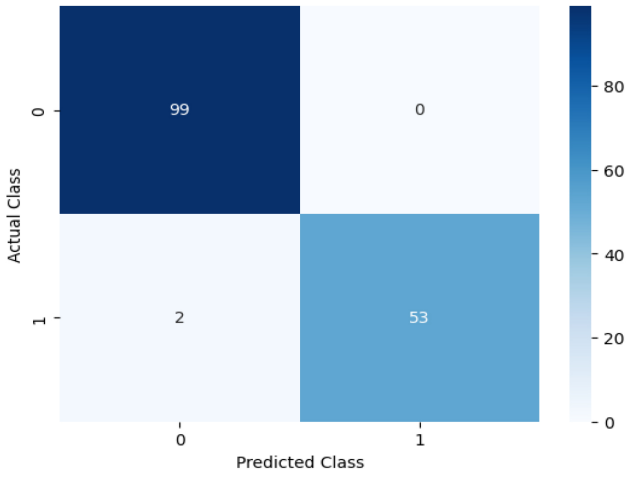

Fig. 12. Confusion matrix of C4.5 algorithm using SMOTE technique.

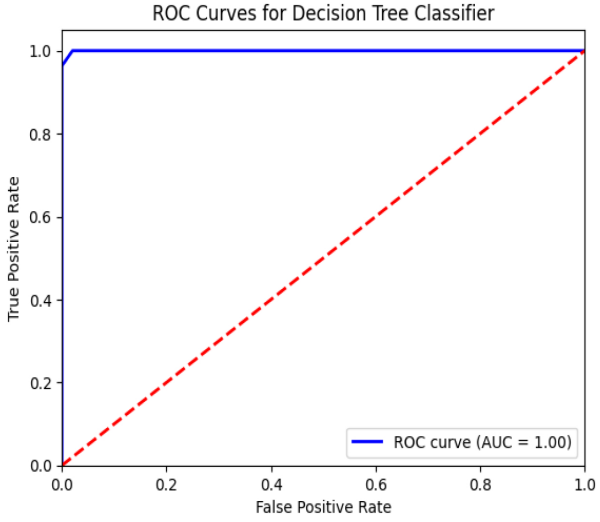

Fig. 13. ROC curve and AUC value for C4.5 with SMOTE.

Fig. 12 shows that 152 cases with TP=99 and TN=53 were accurately identified by the decision tree algorithm using the SMOTE approach.

In Fig. 13, the ROC curve for the C4.5 algorithm using the SMOTE technique is shown. This figure shows that the AUC value for this algorithm is 1.00, which means that the model used achieve ideal performance, i.e. Neither either healthy cases nor any illness cases were mistakenly labeled as negative or positive.

Table 9. Accuracy metrics for the SVM algorithm after using the SMOTE technique.

Class	Precision	Recall	F-Score	Accuracy
0	0.96	1.00	0.98	**97.40%**
1	1.00	0.93	0.96	

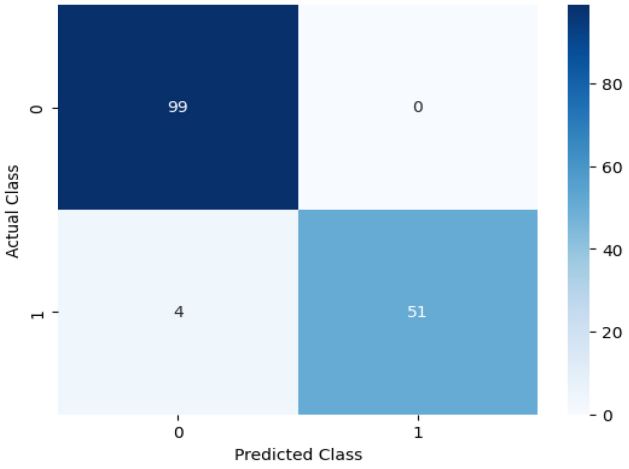

Fig. 14. Confusion matrix of SVM algorithm using SMOTE technique.

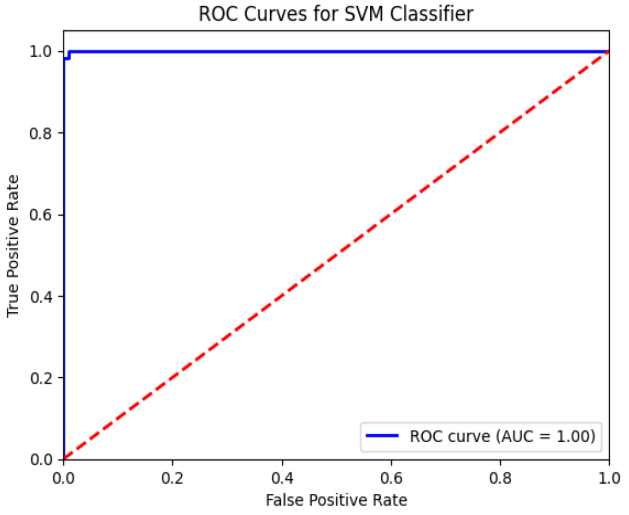

Fig. 15. ROC curve and AUC value for SVM with SMOTE.

Table 9 shows the accuracy metrics for the SVM algorithm after using the SMOTE technique.

While Fig. 14 shows that 150 cases with TP=99 and TN=51 were accurately identified by the SVM algorithm using the SMOTE approach.

In Fig. 15, the ROC curve for the SVM algorithm using the SMOTE technique is shown. This figure shows that the AUC value for this algorithm is 1.00, which means that the model used achieve ideal performance.

Table 10 shows the accuracy measures of the RF algorithm after using this technique.

Table 10. Accuracy measures of the RF algorithm after using the SMOTE technique.

Class	Precision	Recall	F-Score	Accuracy
0	0.99	1.00	0.99	**99.35%**
1	1.00	0.98	0.99	

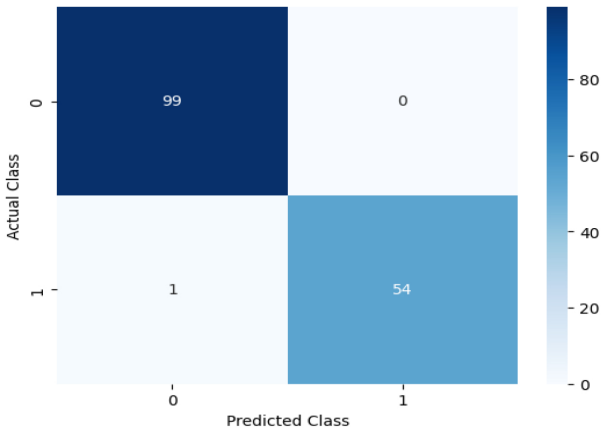

Fig. 16. Confusion matrix of RF algorithm using SMOTE technique.

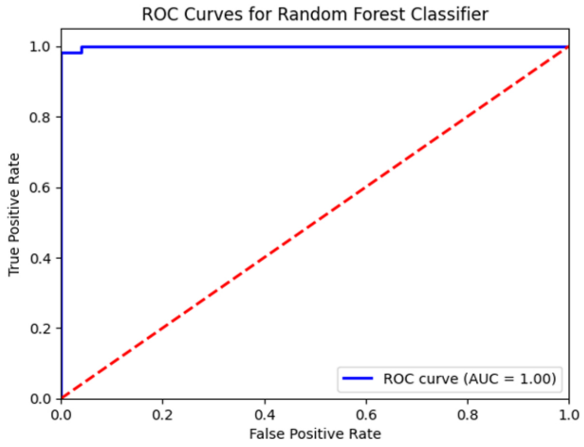

Fig. 17. ROC curve and AUC value for RF with SMOTE.

Fig. 16 shows that 153 cases with TP=99 and TN=54 were accurately identified by the RF algorithm using the SMOTE approach.

In Fig. 17, the ROC curve for the Random Forest algorithm using the SMOTE technique is shown. This figure shows that the AUC value for this algorithm is 1.00, which means that the models used achieve ideal performance, i.e. Neither either healthy cases nor any illness cases were mistakenly labeled as negative or positive.

Table 11 shows the accuracy metrics for the K-NN algorithm after using the SMOTE technique.

Fig. 18 shows that 127 cases with TP=87 and TN=40 were accurately identified by the K-NN algorithm using the SMOTE approach.

Table 11. Accuracy metrics for the K-NN algorithm after using the SMOTE technique.

Class	Precision	Recall	F-Score	Accuracy
0	0.85	0.88	0.87	**82.46%**
1	0.77	0.73	0.75	

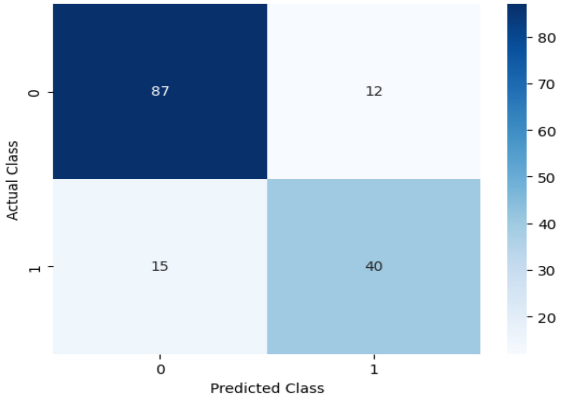

Fig. 18. Confusion matrix of K-NN algorithm using SMOTE technique.

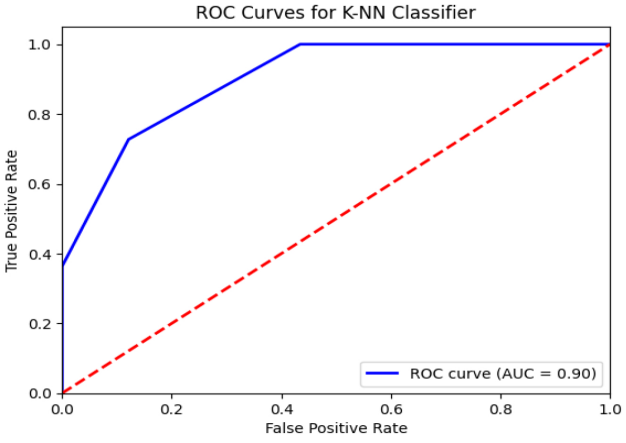

Fig. 19. ROC curve and AUC value for K-NN with SMOTE.

In Fig. 19, the ROC curve for the K-NN algorithm using the SMOTE technique is shown. This figure shows that the AUC value for this algorithm is 0.90, which means that the model used achieve ideal performance.

Table 12 compares the different evaluation measures of the models used of the dataset using the SMOTE synthetic sampling technique.

According to this table, the random forest algorithm reported the best results among the others with 99.35% of accuracy rate and obtained the highest precision rate compared to the rest of the algorithms used with a rate of 99.50 and a recall rate of 99.09 and an F1-score of 99.29, respectively.

Table 12. Evaluation metrics of the models used using SMOTE technique in the dataset.

Algorithms	Accuracy	Precision	Recall	F1-Score
C4.5	98.70	99.00	98.18	98.57
SVM	97.40	98.05	96.36	97.12
RF	99.35	99.50	99.09	99.29
KNN	82.46	81.10	80.30	80.66

7. Conclusion

One of the major issues facing healthcare providers is the early identification of diabetes. Diabetes is challenging to diagnose in its early stages since it can develop gradually without any noticeable symptoms at first. We employed four ML methods in this paper: DT, SVM, RF, and K-NN. Identification for diabetes were developed using the 768 records in the Indian PIMA dataset. The predictive model was tested after eight features were chosen for training. The imbalanced class issue was resolved by using the SMOTE preparation approach. The Random Forest algorithm provided the maximum accuracy of 99.35% for diabetes prediction, according to the trial data. These algorithms may one day be applied to the diagnosis or prognosis of other illnesses.

Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Acknowledgments

Not applicable.

References

Abbas, J. K. K., Ruhaima, A. A., Naser, O. A., & Hayder, D. M. (2024) F-Test and One-Way ANOVA for Medical Images Diagnosis. *Al-Nisour Journal for Medical Sciences*, 6(2), 29–38.

Abdulsahib, A. A., Mahmoud, M. A., Al-Hasnawi, S. A., Almhanna, A. Z., & Abbas, J. K. K. (2025) Automated Retinal Vessel Analysis: A Novel Model for Blood Vessel Detection, Segmentation, and Clinical Characteristic Quantification. *Al-Nisour Journal for Medical Sciences*, 7(1), 65–80.

Acharya, A. (2017) Comparative study of machine learning algorithms for heart disease prediction.

Alrudaini, J. K., Hayder, D. M., Hamzah, A. K., & Ruhaina, A. A. (2022) Visual Perception Method for Medical Image De-noising. Malay. *J. Med. Health Sci*, 18, 40–44.

Bouillon, K., Kivimäki, M., Hamer, M., Shipley, M. J., Akbaraly, T. N., Tabak, A., ... & Batty, G. D. (2013) Diabetes risk factors, diabetes risk algorithms, and the prediction of future frailty: the Whitehall II prospective cohort study. *Journal of the American Medical Directors Association*, 14(11), 851–e1.

Çil, B., Ayyıldız, H., & Tuncer, T. (2020) Discrimination of β-thalassemia and iron deficiency anemia through extreme learning machine and regularized extreme learning machine based decision support system. *Medical hypotheses*, 138, 109611.

Fida, B., Nazir, M., Naveed, N., & Akram, S. (2011) Heart disease classification ensemble optimization using genetic algorithm. In *2011 IEEE 14th International Multitopic Conference* (pp. 19–24). IEEE.

Ibrahim, M. H., & Jaber, A. G. (2022) The Use of the Regression Tree and the Support Vector Machine in the Classification of the Iraqi Stock Exchange for the Period 2019-2020. *Journal of Economics and Administrative Sciences*, 28(132), 74–87.

Islam, M. T., Raihan, M., Farzana, F., Aktar, N., Ghosh, P., & Kabiraj, S. (2020) Typical and non-typical diabetes disease prediction using random forest algorithm. *In 2020 11th International conference on computing, communication and networking technologies* (ICCCNT) (pp. 1–6). IEEE.

Khalifa, M., & Albadawy, M. (2024) Artificial intelligence for diabetes: enhancing prevention, diagnosis, and effective management. *Computer Methods and Programs in Biomedicine Update*, 5, 100141.

Khanam, J. J., & Foo, S. Y. (2021) A comparison of machine learning algorithms for diabetes prediction. *Ict Express*, 7(4), 432–439.

Kumar, M., Hanumanthappa, M., & Kumar, T. S. (2012) Intrusion Detection System using decision tree algorithm. In *2012 IEEE 14th international conference on communication technology* (pp. 629–634). IEEE.

Mienye, I. D., & Jere, N. (2024) A survey of decision trees: Concepts, algorithms, and applications. *IEEE access*.

Mujumdar, A., & Vaidehi, V. (2019) Diabetes prediction using machine learning algorithms. *Procedia Computer Science*, 165, 292–299.

Nahzat, S., & Yağanoğlu, M. (2021) Diabetes prediction using machine learning classification algorithms. *Avrupa Bilim ve Teknoloji Dergisi*, (24), 53–59.

Olisah, C. C., Smith, L., & Smith, M. (2022) Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. *Computer Methods and Programs in Biomedicine*, 220, 106773.

Ovirianti, N. H., Zarlis, M., & Mawengkang, H. (2022) Support Vector Machine Using A Classification Algorithm. *Sinkron: jurnal dan penelitian teknik informatika*, 6(3), 2103–2107.

Panda, M., Mishra, D. P., Patro, S. M., & Salkuti, S. R. (2022) Prediction of diabetes disease using machine learning algorithms. *IAES International Journal of Artificial Intelligence*, 11(1), 284.

Pandey, A., & Jain, A. (2017) Comparative analysis of KNN algorithm using various normalization techniques. *International Journal of Computer Network and Information Security*, 10(11), 36.

Patel, H. H., & Prajapati, P. (2018) Study and analysis of decision tree based classification algorithms. *International Journal of Computer Sciences and Engineering*, 6(10), 74–78.

Patidar, P., Dangra, J., & Rawar, M. K. (2015) Decision tree C4. 5 algorithm and its enhanced approach for educational data mining. *Engineering Universe for Scientific Research and Management*, 7(2), 1–14.

Permana, B. A. C., Ahmad, R., Bahtiar, H., Sudianto, A., & Gunawan, I. (2021) Classification of diabetes disease using decision tree algorithm (C4. 5). *In Journal of Physics: Conference Series* (Vol. 1869, No. 1, p. 012082). IOP Publishing.

Priyam, A., Abhijeeta, G. R., Rathee, A., & Srivastava, S. (2013) Comparative analysis of decision tree classification algorithms. *International Journal of current engineering and technology*, 3(2), 334–337.

Santhanam, T., & Padmavathi, M. S. (2015) Application of K-means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis. *Procedia Computer Science*, 47, 76–83.

Sarwar, M. A., Kamal, N., Hamid, W., & Shah, M. A. (2018) Prediction of diabetes using machine learning algorithms in healthcare. In *2018 24th international conference on automation and computing (ICAC)* (pp. 1–6). IEEE.

Sheng, B., Pushpanathan, K., Guan, Z., Lim, Q. H., Lim, Z. W., Yew, S. M. E., ... & Tham, Y. C. (2024) Artificial intelligence for diabetes care: current and future prospects. *The Lancet Diabetes & Endocrinology*, 12(8), 569–595.

Sheng, B., Pushpanathan, K., Guan, Z., Lim, Q. H., Lim, Z. W., Yew, S. M. E., ... & Tham, Y. C. (2024) Artificial intelligence for diabetes care: current and future prospects. *The Lancet Diabetes & Endocrinology*, 12(8), 569–595.

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

Sonar, P., & JayaMalini, K. (2019) Diabetes prediction using different machine learning approaches. In *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 367–371). IEEE.

Soni, M., & Varma, S. (2020). Diabetes prediction using machine learning techniques. *International Journal of Engineering Research & Technology (IJERT)*, 9(09), 2278–0181.

Thant, A. A., Aye, S. M., & Mandalay, M. (2020). Euclidean, manhattan and minkowski distance methods for clustering algorithms. *International Journal of Scientific Research in Science, Engineering and Technology*, 7(3), 553–559.

Yahyaoui, A., Jamil, A., Rasheed, J., & Yesiltepe, M. (2019) A decision support system for diabetes prediction using machine learning and deep learning techniques. In *2019 1st International informatics and software engineering conference (UBMYK)* (pp. 1–4). IEEE.

Yu, W., Liu, T., Valdez, R., Gwinn, M., & Khoury, M. J. (2010). Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. *BMC medical informatics and decision making*, 10, 1–7.