



A COMPARATIVE EVALUATION OF VARIOUS MACHINE LEARNING TECHNIQUES FOR PREDICTION OF TYPE 2 DIABETES MELLITUS

Rizwan Akhtar^{*1}, Muhammad Kalamuddin Ahamad²

^{1,2}Department of Computer Application, Integral University, Dasauli, Bas-ha Kursi Road, Lucknow, UP, India.

¹<https://orcid.org/0009-0004-7822-4319>^{ID}, ²<https://orcid.org/0000-0003-3043-0549>^{ID}

Email: [*rizwan@iul.ac.in](mailto:rizwan@iul.ac.in), mohdkalam@iul.ac.in

ARTICLE INFO

Article History

Received: March 12, 2026

Reviewed: April 13, 2026

Accepted: July 13, 2026

Published: August 31, 2026

Keywords:

Type 2 Diabetes Mellitus,
Machine Learning,
Predictive Modelling,
Comparative Analysis,
Support Vector Machine,
PIMA Indian Diabetes Dataset,
Clinical Risk Stratification.

ABSTRACT

Diabetes affects over 101 million people in India, with many more at risk due to routine and hereditary factors. Early diagnosis is crucial to prevent complications, which make accurate predictive tools essential in healthcare. This research uses Machine Learning (ML) algorithms to evaluate the likelihood of Type 2 Diabetes Mellitus (T2DM) using lifestyle and family history data. The trained models demonstrate strong predictive ability, allowing individuals to self-assess their risk and supporting healthcare professionals in early detection and intervention. This study presents a performance assessment of seven ML classifiers: Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR), Naïve Bayes (NB), k-Nearest Neighbor (k-NN), and Extreme Gradient Boosting (XGBoost). These classifiers were applied to the widely used PIMA Indian Diabetes dataset (PIDD), which contains 768 clinical records of adult women aged 21 and above, providing key medical information for diabetes analysis. Multiple evaluation measures were applied to assess model performance with results showing that SVM achieved the highest accuracy and AUC, while LR, RF, and XGBoost also performed competitively. Although k-NN attained the highest recall, it yielded a higher false positive rate. These findings highlight that no single model is perfect for every situation, and the choice of classifier should match clinical needs. This study serves as a reference for ML applications in diabetes prediction.



Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Diabetes mellitus (DM) is a chronic endocrine-metabolic condition marked by abnormal regulation of blood glucose levels due to the body's inability to produce sufficient insulin, improper utilization of insulin, or both. When insulin activity is disrupted, glucose accumulates in the bloodstream, impairing the normal metabolism of carbohydrates, lipids, and proteins [1]. Chronic high blood glucose interferes with the metabolism of carbohydrates, fats, and proteins. These metabolic imbalances can result in serious long-standing complications, impacting organs including the eyes, kidneys, nerves, and the cardiovascular system [2]. As a progressive disease, diabetes requires timely diagnosis, continuous monitoring, and effective management to reduce its significant health and economic burden [3]. Inadequate management can lead to acute conditions like diabetic ketoacidosis. Diabetes is a leading cause of significant morbidity, including cardiovascular disease, stroke, renal failure, diabetic foot ulcers, and retinopathy [4]. The significant health burden and mortality resulting from these complications highlight the imperative for early detection and effective long-term management [5]. Based on their underlying causes, pathophysiological mechanisms, and treatment considerations, diabetes is primarily divided into three main forms: Type 1, Type 2, and Gestational Diabetes Mellitus (GDM) [6]. Each type differs in its origin, clinical presentation, and management approach. Type 1 Diabetes Mellitus (T1DM) results from autoimmune destruction of pancreatic beta-cell, requires patients to rely on external insulin administration for survival. Type 2 Diabetes Mellitus (T2DM) occurs from impaired insulin action and insufficient insulin production commonly associated with obesity and physical inactivity. Gestational diabetes mellitus (GDM) is a condition of pregnancy that confers health risks to both the mother and infant.

For a healthy individual, normal fasting blood glucose is 70-99 mg/dL, a level of 100-125 mg/dL shows prediabetes, and exceeding 126 mg/dL diagnoses diabetes are classified as pre-diabetic [7]. Such a person is prone to the development of T2DM. Research has identified several factors that contribute to a higher likelihood of developing diabetes.

- (a) Weight-to-height ratio exceeding 25
- (b) A history of diabetes in close relative
- (c) Low HDL cholesterol levels (less than 40 mg/dL)
- (d) Long-standing hypertension or history of gestational diabetes
- (e) Previous diagnosis of polycystic ovary syndrome (PCOS)
- (f) Key non-modifiable risk factors for diabetes include age over 45 years, family history, and belonging to high-risk ethnic groups.
- (g) Sedentary or inactive lifestyle.

Management of prediabetes centers on lifestyle interventions designed to prevent or delay the development of T2DM [8]. This research focuses on evaluating the risk of T2DM development in at-risk individuals. The organization of paper is as follows: section II outlines the existing body of related research, while section III described the classification techniques applied. The performance metrics for evaluating the classifiers are detailed in section IV. Section V outlines the methodological framework and implementation procedures, followed by the experimental configuration in section VI. The results and their analysis are presented in Section VII. Finally, Section VIII concludes the paper with a synthesis of the findings and future research avenues.

II. RELATED WORKS

This section summarizes important research on predicting diabetes, focusing on the application of various ML algorithms and comparing different classification methods. The rising global rate of T2DM has drawn considerable attention to its prediction, with data-driven methods offering valuable support for early detection and prevention [8]. Predicting T2DM mainly involves supervised learning, which has led researchers to frequently use ML and data mining techniques (DMT). While the PIDD is a standard benchmark, recent research has shifted the focus beyond just accuracy. Current efforts aim to improve interpretability, tackle class imbalance, and enhance early detection through methods like artificial neural networks and genetic algorithms. Research identified physical activity as a critical factor in mitigating T2DM risk in prediabetic individuals, underscoring the necessity of incorporating lifestyle variables into predictive models [9]. Studies comparing ML techniques for early-stage diabetes prediction have shown that ensemble methods frequently achieve better performance than traditional classifiers [10]. Similarly, oversampling strategies were used to improve early detection of female T2DM cases, addressing issues with imbalanced datasets [11]. Additionally, several studies proposed novel frameworks for diabetes classification. Machine learning was applied to classify T2DM subtypes, allowing more personalized interventions [12]. A hybrid meta-classifier combining fuzzy clustering with LR showed higher accuracy compared to traditional models [13]. Earlier work using DMT also contributed significantly to establishing prediction models for T2DM [14]. Recent research has focused on ensemble and hybrid learning to enhance predictive capabilities. For instance, a tri-ensemble model using KNN-imputed features significantly improved prediction outcomes [15]. Systematic evaluation of ML algorithms for prediction of diabetes has established critical benchmarks for evaluating their performance and computational efficiency [16]. This study introduced correlation based model that leverages both machine and deep learning (DL) for automatic prediction of diabetes, demonstrating improved effectiveness compared to traditional methods [17]. Another study proposed a risk assessment framework that integrates multiple machine learning algorithms, making them suitable for real-world healthcare applications [18].

Table 1: Summarized related works from literature.

SN	References	Methodology	Research Focus	Key Findings	Limitations
1	Afolabi, S., et al. [19]	Supervised ML algorithms (e.g., SVM, DT)	Imbalanced datasets, missing values	Achieved high accuracy in diabetes prediction	Limited generalizability across diverse populations
2	Kim, J., et al. [20]	Deep learning model using ECG data	Long-term risk assessment	Identified long-term diabetes risk using ECG	Requires large datasets for training
3	Mohsen, F., et al. [21]	Multimodal deep learning model integrating ECG features	Lack of integration of ECG in traditional models	Enhanced prediction accuracy with ECG data	Potential overfitting with complex models
4	Kulkarni, A. R., et al. [22]	ML on ECG signals	Non-invasive early detection	Achieved 80% accuracy in pre-diabetes detection	Limited to specific demographic groups
5	Lee, H., et al. [23]	AI-based prediction model	Aging population with comorbidities	Improved prediction accuracy in older adults	Requires validation across different age groups
6	Abnoosian, K., et al. [24]	Ensemble machine learning (e.g., Random Forest, SVM)	Classification of diabetes stages	High accuracy in multi-class classification	Computationally intensive
7	Ansari, Y., et al. [25]	Deep learning architectures (e.g., CNN, RNN)	Arrhythmia detection and diabetes prediction	Achieved high performance in both tasks	Requires large labeled datasets
8	Alzboon, M. S., et al. [26]	Comparison of ML algorithms (e.g., Logistic Regression, SVM)	Early detection of diabetes	Neural Network outperformed other models	Limited interpretability of complex models
9	Khokhar, P. B., et al. [27]	Machine learning with explainable AI (e.g., SHAP, LIME)	Model interpretability	Achieved 92.5% accuracy with high interpretability	Potential bias in feature selection
10	Hennebelle, A., et al. [28]	IoT-edge-AI-cloud integrated system	Real-time diabetes prediction	Improved prediction with integrated system	System complexity and integration challenges

Source: Authors, (2026).

The literature shows the clear shifts in diabetes research from traditional classifiers to advanced ensemble and DL models, driven by the integration of diverse data-driven methods to improve T2DM prediction and management.

However, as summarized in Tables 1 and 2, no single algorithm demonstrates universal superiority. Empirical performance is highly contingent on specific contextual factors, such as dataset structure and size. Thus, the choice between models like SVMs and neural networks is problem-dependent.

II.1 KEY CONTRIBUTION OF RESEARCH WORK

Major key contribution of this study focuses on effectiveness of various ML algorithms for predicting t2dm. The study also highlights comparative performance analysis and also identifies most suitable model for accurate and reliable prediction.

- (a) A comparative framework evaluates seven ML classifiers for T2DM prediction.
- (b) Performance assessed by multiple metrics, ensuring balanced evaluation beyond simple accuracy measures.
- (c) The SVM classifier outperformed other models, attaining optimal accuracy and AUC metrics.
- (d) Trade-offs highlighted, showing models vary between sensitivity, specificity, and false positive rates.
- (e) Study establishes a benchmark framework for applying machine learning in real-world diabetes prediction tasks.

III. OVERVIEW OF CLASSIFICATION TECHNIQUES

This section outlines the key features of the seven individual and ensemble classifiers examined in this research.

III.1 DECISION TREE

A DT is a simple yet powerful model used in machine learning and decision analysis to make predictions or choices by splitting data into branches. DT is a classification technique that organizes decision making in hierarchical tree like structure.

Table 2: Classification accuracies reported in various research.

SN	References	Technique Used	Accuracy	Dataset Used
1	Jena et al.[29]	LR, DT, RF, SVM, and GB as ensemble	84.2	PIMA Indians
2	Belqes alsadi et al.[30]	XG Boost, RF	84	Qatar Bio Bank(1000 Participants)
3	Rastogi et al.[31]	RF,SVM,LR,NB	LR-82.46,SVM-79.22,NB-79.22, RF-81.81	PIMA Indians
4	Agliata et al.[32]	ANN	86	NHANES,MIMIC-III,MMIC-IV
5	Roobini et al.[33]	Adaboost	83	PIMA Indians
6	Khanam et al.[34]	NN	88.6	PIMA Indians
7	Joshi et al.[35]	LR, DT	78.26	PIMA Indians
8	Tigga et al.[36]	RF, SVM, DT, kNN, LR, NB	LR-85,74,KNN-77,70,SVM-86,74 NB-80,68,DT-84,69 RF-94,72	952 Questannaire+PIMA
9	Alam et al. [37]	ANN,RF,K-Means clustering	RF-74,ANN-75,K Means-73	National Institute of Diabetes and Digestive and Kidney Diseases
10	Mirzajani et al. [38]	DT, NN, Bayesian Network, SVM	DT-80.2,NN-77.6,BN-77.8, SVM-72.73	PIMA Indians

Source, Authors, (2026).

Internal nodes indicate tests on patient attributes such as glucose level, BMI, or age, leading through branches to leaf nodes that shows the final outcome—diabetic or non-diabetic [39]. DT is favored in healthcare for their interpretable rules, minimal preprocessing, and versatility with data types. In the context of T2DM prediction, DT shown in fig 1 can help clinicians in identifying key risk factors and forming diagnostic rules but may overfit small or complex datasets, reducing accuracy in multi-class or imbalanced cases [40]. The pseudocode for decision tree is given in figure 2.

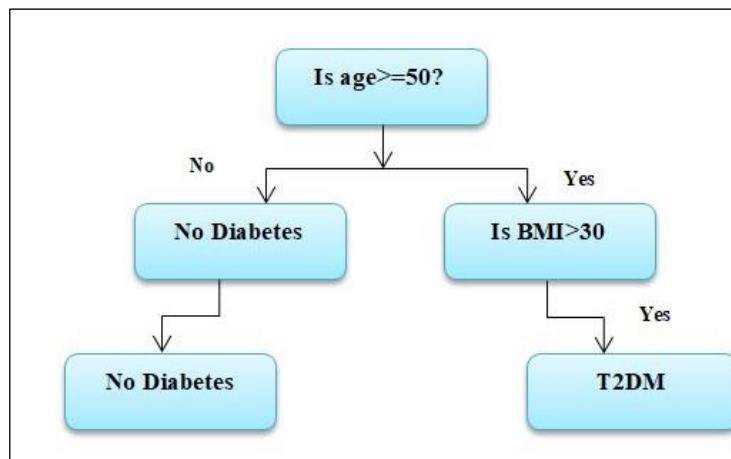


Figure 1: Decision Making in T2DM Prediction.

Source: Authors, (2026).

Input: Training Data D with features and class labels

1. If all records in D belong to the single class
→ Assign that class to the leaf node
2. Else if no feature remains for splitting
→ Return the majority class in D (Leaf Node)
3. Else
 - 3.1. Select the best splitting feature A (using Information Gain / Gini Index)
 - 3.2. Create a decision node for feature A
4. For each possible value v of feature A :
 - 4.1. Partition dataset D_v where $A = v$
 - 4.2. If D_v is empty → assign majority class of D
 - 4.3. Else → recursively apply Decision Tree on D_v

Output: Constructed Decision Tree for classification

Figure 2: Pseudocode of decision tree.
Source: Authors, (2026).

III.2 RANDOM FOREST (RF)

RF classifier operates as an ensemble approach that leverages bagging and random feature selection to build numerous decorrelated decision trees. This structure promotes high generalization by aggregating their predictions via majority voting [41]. Due to its robustness and capacity to manage diverse types of data, RF has been widely utilized for predicting T2DM. Working of Random Forest is shown in fig 3 and pseudocode for RF is given in figure 4.

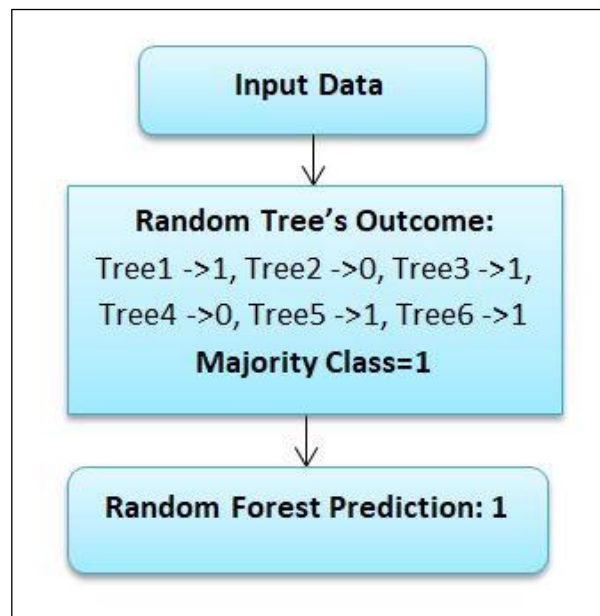


Figure 3: Workflow of RF Algorithm.
Source: Authors, (2026).

Input: Dataset Training D with patient samples n and features m .

1 Bootstrap Sampling:

Generate k bootstrapped datasets $D_1, D_2, \dots, D_k$ from D , each obtained by sampling with replacement.

2 Tree Construction:

For each dataset D_i :

- 2.1 Build a decision tree by recursively splitting nodes.

2.2 At each split, randomly select a subset of features $F \subseteq m$.

2.3 Choose the best feature from F based on a splitting criterion.

3 Model Training:

3.1 Repeat Step 2 until N independent decision trees are constructed.

4 Prediction (for a test instance x):

4.1 Each tree outputs a predicted label y_i .

4.2 Aggregate the predictions using majority voting:

$$\hat{y} = \operatorname{argmax}_{c \in \{0,1\}} \sum_{i=1}^N I(y_i = c)$$

where, $I(\cdot)$ denotes the indicator function.

Output: Final predicted class $\hat{y} \in \{\text{Diabetes, No Diabetes}\}$.

Figure 4: Pseudocode of Random Forest.

Source: Authors, (2026).

III.3 SUPPORT VECTOR MACHINE (SVM)

SVM is a supervised classification method that classifies data by creating an optimal hyperplane to maximally separate distinct classes. It performs effectively across diverse data patterns and complex feature distributions [42]. SVM aims to maximize the separation margin between classes, which improves generalization. The margin is set by the support vectors, which are the key points closest to the decision boundary. For linearly separable data, the hyperplane is defined by: $w^T x + b = 0$, where $w = \{w_1, w_2, w_3, \dots, w_n\}$ denotes the weight vector, b is the bias term, and n is the number of attributes. The classification constraints can be written as:

$$\begin{aligned} \text{(i)} \quad & w^T x_i + b \geq 1, \quad \text{if } y_i = +1 \\ \text{(ii)} \quad & w^T x_i + b \leq -1, \quad \text{if } y_i = -1 \end{aligned}$$

SVM identifies the hyperplane that maximizes the distance to the nearest training points. This principle is captured by the following optimization formulation:

$$\text{Margin} = \frac{2}{\|w\|}$$

For non-linear problems, kernel functions like Polynomial, RBF, and Sigmoid handle non-linear data by projecting it into higher dimensions, facilitating linear separation. The classification decision is then determined by the resulting function:

$$f(x) = \operatorname{sign}\left(\sum_{i=1}^l \alpha_i y_i K(x_i, x) + b_0\right)$$

Where y_i represents the class label of support vector x_i , α_i are lagrange multipliers, $K(x_i, x)$ denotes the kernel function, b_0 is the bias term, and l indicates the numbers of support vectors. Working of SVM is shown in figure 5 and pseudocode for SVM is given in figure 6.

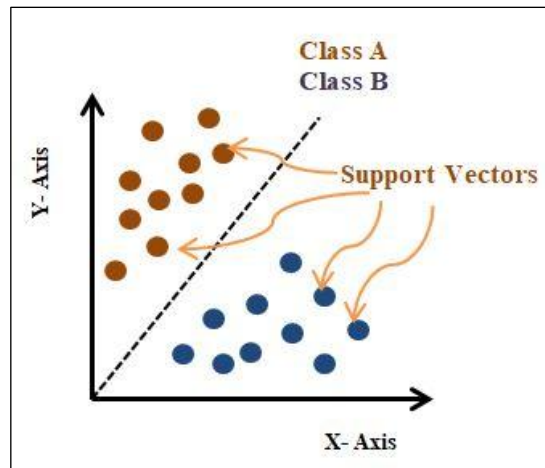


Figure 5: Support Vector Machine.

Source: Authors, (2026).

Input: Training Data, Test Instance

1. Choose the value of kernel function $K(x_i, x_j)$ (Linear, Polynomial, RBF, Sigmoid).
2. Formulate the optimization problem to find the hyperplane: $w^T x + b = 0$
3. Identify support vectors that lie nearest to the hyperplane.
4. Construct the decision function:

$$f(x) = \text{sign}\left(\sum_{i=1}^l \alpha_i y_i K(x_i, x) + b_0\right)$$

5. For the given test instance:

- 5.1 If $(x) \geq 0 \Rightarrow$ assign to class +1 (Diabetes)
- 5.2 Else $\Rightarrow$ assign to class -1 (No Diabetes)

Output: Predicted Class (Diabetes / No Diabetes)

Figure 6: Pseudocode of SVM.
Source: Authors, (2026).

III.4 LOGISTIC REGRESSION(LR)

This statistical technique is widely used for binary classification, estimating the probability that an observation belongs to one of two categories rather than generating continuous numerical outputs [43]. The model assumes that the log-odds of the probability are a linear function of the input features. For an input feature vector $X = (x_1, x_2, x_3, \dots, x_n)$, the LR techniques is denoted as:

$$p(x) = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + b_2 x_2 + \dots + b_n x_n)}}$$

Where, b_0 is the intercept, $b_1, b_2, \dots, b_n$ are regression coefficients, and n is the number of input features. The decision boundary is defined at $p(x) = 0.5$. Thus, classification can be expressed as:

$$\hat{y} = \begin{cases} 1 & \text{if } p(x) \geq 0.5 \quad (\text{Diabetes}) \\ 0 & \text{if } p(x) < 0.5 \quad (\text{No Diabetes}) \end{cases}$$

Model parameters are optimized using Maximum Likelihood Estimation by optimizing the cost function to find the best-fitting coefficients. LR working is shown in figure 7 and pseudocode in figure 8.

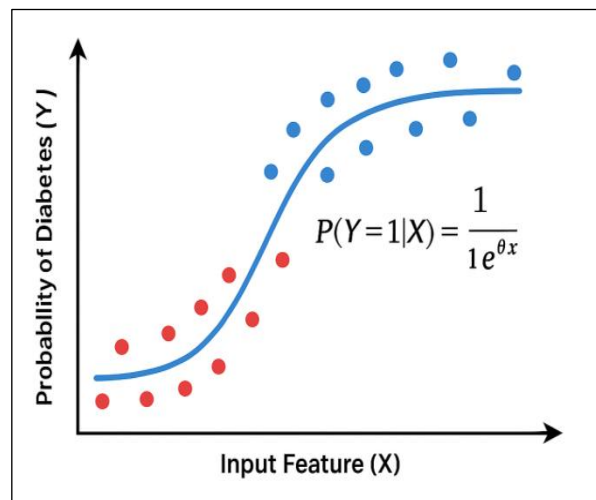


Figure 7: Workflow of Logistic Regression.
Source: Authors, (2026).

Input: Training Data, Test Instance

1. Assign starting value weights w and bias b .
2. For every training instance:
 - 2.1 Compute probability using sigmoid:

$$p(x) = \frac{1}{1+e^{-(w^T x + b)}}$$

2.2 Modify weights w and bias b using gradient descent.

3. Repeat step 2 until convergence.
4. For the given test instance:
 - 4.1 Compute probability $p(x)$.
 - 4.2 If $p(x) \geq 0.5 \Rightarrow$ assign to class +1 (Diabetes).
 - 4.3 Else $\Rightarrow$ assign to class 0 (No Diabetes).

Output: Predicted Class (Diabetes / No Diabetes)

Figure 8: Pseudocode of Logistic Regression.
Source: Authors, (2026).

III.5 NAÏVE BAYES (NB)

The NB classification method applies ‘Bayes Theorem’ treating each feature as independent from the others. Its simplicity and efficiency make it effective, especially for high-dimensional datasets. [44]. Working of NB is shown in figure 9 and the pseudocode is given in figure 10.

Model Representation:

$$P(C_k|x) = \frac{P(x|C_k) P(C_k)}{P(x)}$$

Where:

$P(C_k|x)$: posterior probability of class C_k given features

$xP(x|C_k)$: probability of observing x given class

$C_k P(C_k)$: prior probability of class

$C_k P(x)$: evidence (same for all classes, so ignored in decision)

Using the naïve assumption of conditional independence:

$$P(x|C_k) = \prod_{i=1}^n P(x_i|C_k)$$

Final classification rule:

$$\hat{y} = \underset{C_k}{\operatorname{argmax}} P(C_k) \prod_{i=1}^n P(x_i|C_k)$$

Decision Rule:

$$\hat{y} = \begin{cases} 1 & \text{if posterior probability is maximum for Diabetes} \\ 0 & \text{if posterior probability is maximum for No Diabetes} \end{cases}$$

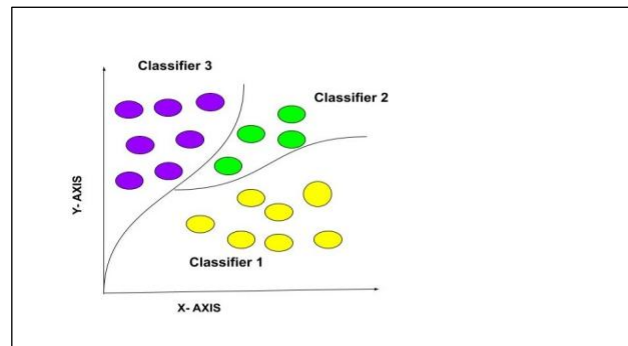


Figure 9: Naïve Bayes Classifier.
Source: Authors, (2026).

Input: Training Data , Test Instance

Compute prior probability of every class:

$$P(C_k) = \frac{\text{Samples count in class } C_k}{N}$$

For each feature x_i :

- a. Calculate $P(x_i|C_k)$ for all possible classes.
- b. (For continuous features, assume Gaussian distribution and compute using probability density function.)

For a new instance $x = (x_1, x_2, \dots, x_n)$:

- c. Compute posterior probability for each class:

$$P(C_k|x) \propto P(C_k) \prod_{i=1}^n P(x_i|C_k)$$

Assign the observation to the class exhibiting the maximum.

Output: Predicted Class (Diabetes / No Diabetes).

Figure 10: Pseudocode of Naïve Bayes.
Source: Authors, (2026).

III.6 K-NEAREST NEIGHBOR (KNN) CLASSIFIER

k-NN is a simple supervised technique that predicts a new instance's class based on the k nearest neighbors, without explicit training [45]. Closeness is typically measured using Euclidean distance, with other metrics applied as needed [46]. Figure 11 shows how the value of k influences classification in the k-NN algorithm and the pseudocode of k-NN is given in figure 12.

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

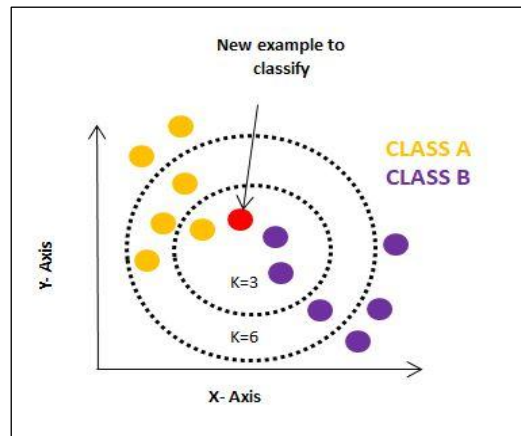


Figure 11: kNN Classifier.
Source: Authors, (2026).

Input: Training Data, Test Instance

1. Choose the appropriate value of k
2. For every instance i in the training dataset:
 - 2.1 Calculate the distance D_i between the test instance with instance i
3. Sort all computed distances D_i in increasing order of distance
4. Choose the k nearest neighbors from the sorted list
5. Determine the most frequent class label among these neighbors
6. Assign this class label to the test instance

Output: Predicted Class (Diabetes / No Diabetes)

Figure 12: Pseudocode of k-Nearest Neighbor (k-NN).
Source: Authors, (2026).

III.7 EXTREME GRADIENT BOOSTING (XGBOOST) CLASSIFIER

XGBoost is a refined form of gradient boosting that constructs DT sequentially, with each tree correcting prior errors. It employs regularization, shrinkage, and feature subsampling, enhancing robustness and performance on structured data such as diabetes prediction. [47]. XGBoost minimizes classification errors by optimizing a loss function.

Each tree fits the residuals of the previous ones, using gradients and Hessians to maximize information gain, and combines all trees for the final prediction of T2DM. XGBoost prediction process is shown in figure 13 and the pseudocode for that is shown in figure 14.

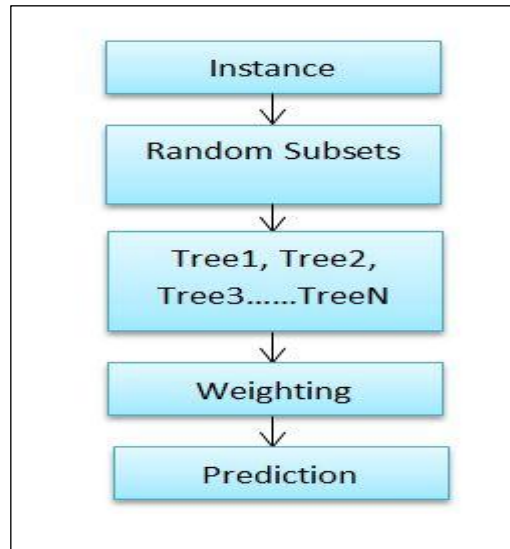


Figure 13: XGBoost prediction process.
Source: Authors, (2026).

Input: Training Dataset, Number of Trees (T), Learning Rate (η), Maximum Depth, Regularization Parameters

1. Initialize the model with a base prediction (e.g., mean of labels).
2. For every boosting round $t = 1$ to T :
 - 2.1 Compute the gradient (g_i) and hessian (h_i) of the loss function for each training instance.
 - 2.2 Construct a DT to fit the residuals:

- For each feature and split point:
 - Calculate the information gain using g_i and h_i .
 - Choose the split that maximizes gain.

2.3 Apply regularization to prune and avoid overfitting.

2.4 Update the prediction for each instance using:

$$\hat{y}^{(t)} = \hat{y}^{(t-1)} + \eta \cdot f_t(x)$$

where, $f_t(x)$ is the output of the newly added tree.

3. After all trees are built, combine their weighted outputs to form the final prediction.

4. Assign the predicted class label based on threshold (e.g., $\geq 0.5 \rightarrow$ Diabetic, otherwise Non-Diabetic).

Output: Predicted Class (Diabetes / No Diabetes)

Figure 14: Pseudocode of Extreme Gradient Boosting (XGBoost) Classifier.
Source: Authors, (2026).

IV. OVERVIEW OF CLASSIFICATION TECHNIQUES

Evaluating a machine learning classifier involves defining suitable performance metrics that measure how effectively the model distinguishes between classes. This study uses nine such metrics, which are briefly described below:

Confusion Matrix (CM): A structured table used to show classification accuracy, with rows representing actual class labels and columns displaying predicted labels. Correctly classified instances lie along the diagonal of the matrix.

True Positive (TP): Instances in which the classifier correctly identifies positive cases as positive.

False Positive (FP): Observations incorrectly predicted as when they are actually negative.

True Negative (TN): Instances correctly classified by the model as negative.

False Negative (FN): Cases where positive instances are mistakenly classified as negative.

Based on these four quantities, multiple evaluation measures can be formulated:

IV.1 CLASSIFICATION ACCURACY

The number of correctly classified examples divided by the total number of observations.

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

IV.2 PRECISION

Precision indicates the fraction of correctly predicted positive cases among all instances predicted as positive.

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

IV.3 SPECIFICITY

Specificity represents the fraction of actual negative cases that is correctly identifies.

$$Specificity = \frac{TN}{TN + FP}$$

IV.4 RECALL/ SENSITIVITY/ TRUE POSITIVE RATE

Recall evaluates the actual positives correctly detected.

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

IV.5 FALSE POSITIVE RATE (FPR)

This measure calculates the fraction of negative cases that are incorrectly predicted as positives.

$$FPR = \frac{FP}{FP + TN}$$

IV.6 NEGATIVE PREDICTIVE VALUE

NPV is the proportion of negative prediction that are accurate among all predicted negatives.

$$NPV = \frac{TN}{TN + FN}$$

IV.7 RATE OF MISCLASSIFICATION (RMC)

RMC is the proportion of incorrect predictions, including Type-I (positive as negative) and Type-II (negative as positive) errors.

$$RMC = \frac{FP + FN}{TP + TN + FP + FN}$$

IV.8 F1-SCORE

It is Harmonic mean of precision and recall, useful for imbalanced data.

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

IV.9 RECEIVER OPERATING CHARACTERISTIC (ROC) CURVE

The ROC curve evaluates the classifier effectiveness by plotting the trade-off between sensitivity and specificity at multiple thresholds. Greater AUC values represent better capability to distinguish between classes.

V. DESIGN AND IMPLEMENTATION

This research analyzed multiple ML models for prediction of T2DM using the PIDD. After careful data preparation and an 80-20 stratified split, we evaluated different techniques, including DT, RF, LR, SVM, kNN, NB, and XGBoost. We developed the models using k-fold cross-validation to improve their settings. We assessed performance using a separate test set and multiple metrics. The comparative analysis found the best technique based on its ability to generalize and its usefulness in clinical settings. This provided clear evidence for choosing the right algorithm for diabetes prediction. We compared the performance of these classifiers systematically, and Figure 15 illustrates the process for implementing the algorithms.

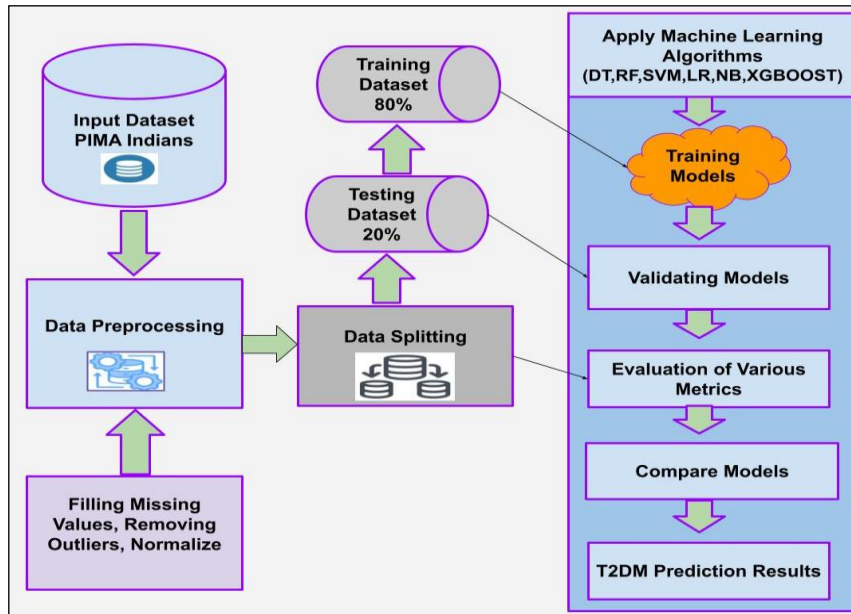


Figure 15: Flowchart for T2DM prediction.

Source: Authors, (2026).

To ensure reproducibility and clarity of the experimental procedure, the adopted workflow is presented in the form of an algorithm Figure 16.

Input: PIMA Indian Diabetes Dataset (with attributes such as Glucose, BMI, Age, Insulin, etc.)

Step 1: Data Collection & Preprocessing

1.1 Import dataset.

1.2 Handle missing or zero values

1.3 Normalize/standardize continuous features

Step 2: Data Splitting

2.1 Divide dataset Training (80%) and Testing (20%)

Step 3: Model Training

Seven Machine Learning Classifiers are chosen for comparison

$m = [kNN(), LR(), SVM(), NB(), DT(), RF(), XGB()]$

for ($k=0$; $k<7$; $k++$) do

Model= $m[i]$;

Model.fit();

Model.predict();

print(Accuracy(i), confusion_matrix, classification_report, roc_curve)

End

Step 4: Model Evaluation

4.1 Test classifiers on unseen test data

4.2 Generate confusion matrix for each model

4.3 Compute evaluation metrics

Step 5: Comparative Analysis

5.1 Compare results of all classifiers

5.2 Identify best-performing model based basis on Performance Metrics

Output: Predicted Class (Diabetes / No Diabetes)

Figure 16: Algorithm for Prediction of T2DM.

Source: Authors, (2026).

The pseudocode outlines sequential steps: data collection and preprocessing, model selection, training, and evaluation. Classifiers were trained on processed data and tested on unseen instances, with performance assessed using standard metrics, followed by a comparative analysis to determine the optimal model for diabetes risk prediction.

VI. EXPERIMENTAL SETUP

An empirical analysis of diabetes prediction was conducted using the PIDD. Classifiers including DT, RF, XGBoost, SVM, k-NN, LR, and NB were evaluated using performance metrics. Experiments took place in Python with scikit-learn, pandas, numpy, and matplotlib within Jupyter Notebook and Google Colab for reproducibility and visualization. The workflow treated datasets as inputs, applied classifiers, and assessed the results with the Test & Score tool shown in figure 17.

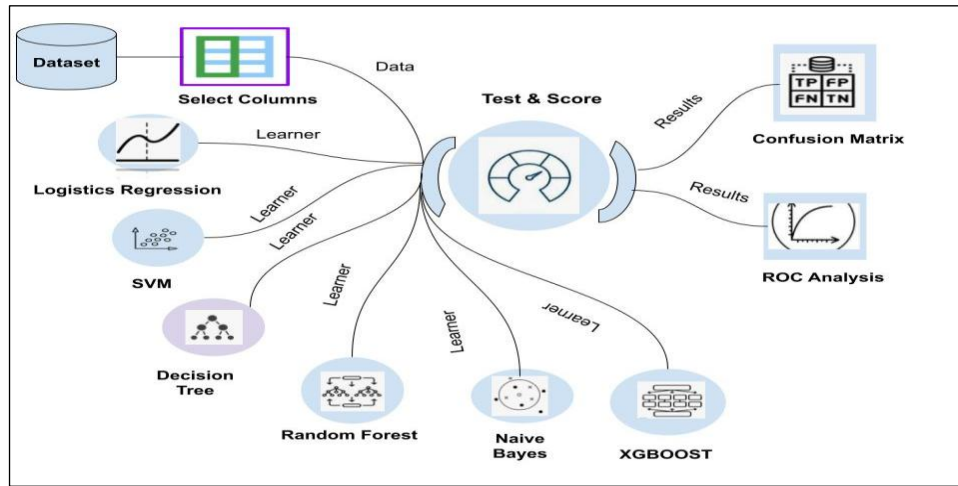


Figure 17: Experimental setup.

Source: Authors, (2026).

The outcomes were further visualized using confusion matrix, ROC curve, and evaluation metrics.

VI.1 DATASET DESCRIPTION

The PIDD comes from the UCI Machine Learning Repository. It includes data on 768 female patients of Pima Indian heritage, aged 21 and older. The dataset has 8 medical and physiological features, along with a variable that shows whether diabetes is present [47]. Table 3 lists the features of the dataset.

Table 3: Dataset Features Information.

SN	Attributes	Abbreviation	Description
1	Pregnancy Count	Preg	Count of pregnancies
2	Plasma glucose concentration	Glu	Blood glucose level after 2-hour test
3	Diastolic blood pressure (mm Hg)	BP	Patient's diastolic blood pressure reading
4	Triceps skinfold thickness (mm)	Skin	Measurement of skin fold thickness
5	Two-Hour serum insulin (mu U/ml)	Insulin	Insulin concentration in the blood
6	Body Mass Index (BMI)	BMI	Weight in kg / Height in m ²
7	Diabetes pedigree function	DPF	Score representing the patient's hereditary risk of diabetes
8	Age (years)	Age	Age of Individual
9	Class variable	Class	0 = non-diabetic, 1 = diabetic

Source: Authors, (2026).

The correlation matrix of the PIDD, figure 18 indicates that Glucose shows the highest positive association with diabetes outcome (0.47). This is followed by BMI (0.29), Age (0.24), and Preg (0.22). Other features like BP, Skin Thickness, Insulin, and DPF exhibit weaker correlations, suggesting limited predictive value on their own. There are moderate correlations between features, such as Glucose and Insulin (0.33) and BMI and Skin Thickness (0.39), show some redundancy among the variables [48].

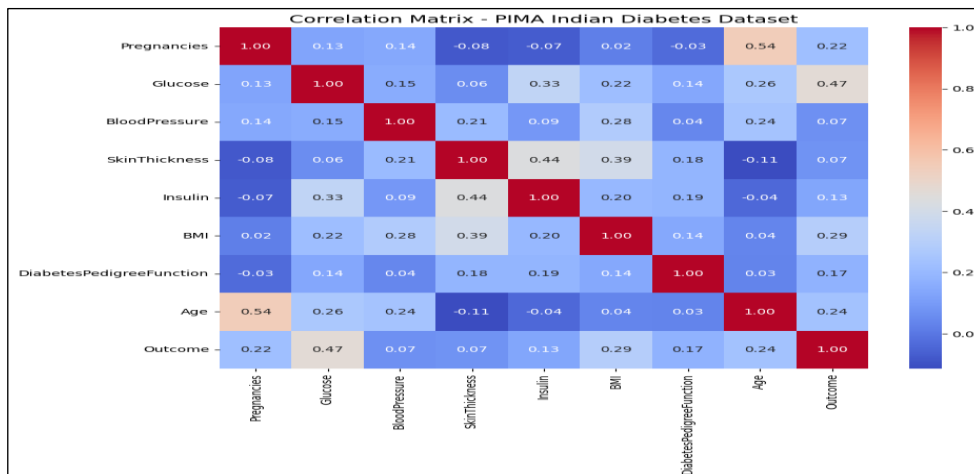


Figure 18: Correlation Matrix – PIMA Indian Diabetes Dataset (PIDD).

Source: Authors, (2026).

Overall, the analysis highlights that while a few variables are moderately linked to diabetes, most exhibit weak linear relationships, underscoring the need for ML models to capture complex, non-linear patterns.

VII. RESULT AND ANALYSIS

This research evaluated the performance of seven predictive models was assessed for predicting T2DM. The evaluation used multiple performance metrics. Table 11 summarizes the classification results, while Figure 19 illustrates the ROC curves for comparative analysis and figure 20 shows performance statistics of seven classifiers. The SVM model outperformed others, recording an AUC of 0.86 and a classification accuracy of 0.90. It also showed strong precision at 0.93 and specificity at 0.97, indicating its reliability in identifying non-diabetic cases. **Logistic Regression** had comparable results with an AUC of 0.83 and a classification accuracy of 0.88, excelling in precision at 0.92 and specificity at 0.97. However, recall was slightly lower at 0.69, suggesting a higher chance of missing positive (diabetic) cases. **Random Forest** and **XGBoost** performed competitively, with AUC scores of 0.84 and 0.85, respectively. **Random Forest** achieved higher precision at 0.93, but had slightly lower recall at 0.71. **XGBoost** maintained equilibrium between precision and recall, attaining 0.83 and 0.80, respectively. Both ensemble methods showed robustness and had lower false positive rates ($FPR \leq 0.09$), making them suitable for medical decision-making, where minimizing false alarms is important. **Naïve Bayes** had moderate results with an AUC of 0.83 and a classification accuracy of 0.86. It performed well in terms of specificity at 0.93, but slightly underperformed in recall at 0.72. This indicates that the model effectively identified non-diabetic individuals but was less accurate in detecting those with diabetes. The **k-NN** classifier exhibited the highest recall at 0.90, indicating strong sensitivity to positive cases, which is desirable in medical diagnosis. However, its relatively low precision at 0.69 and high false positive rate at 0.19 suggest that it tends to over-predict diabetes, potentially causing unnecessary alarms. The **Decision Tree** model, although simple, produced competitive outcomes with an AUC of 0.84 and a classification accuracy of 0.86. It provided balanced precision and recall, both at 0.79, though it underperformed compared to ensemble-based methods in terms of robustness.

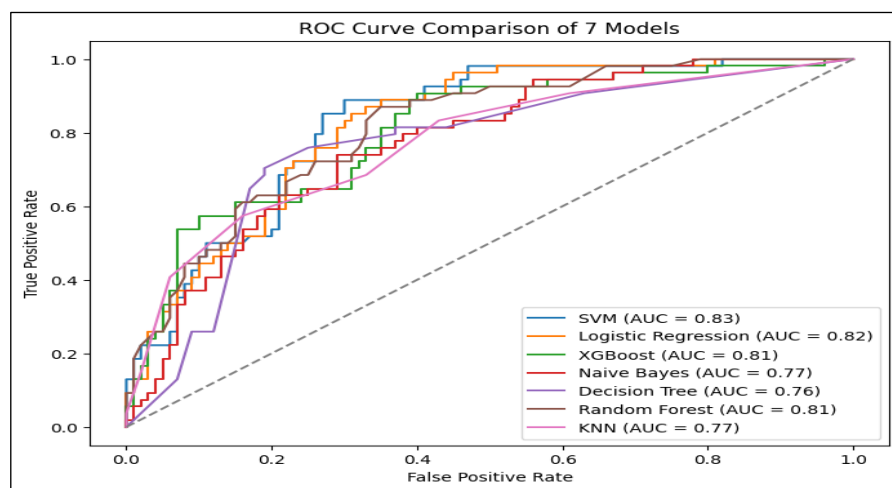


Figure 19: ROC Curve of seven classifiers.
Source: Authors, (2026).

The **ROC curve** comparison figure 16 reinforces these findings. SVM and Logistic Regression achieved the most consistent separation between classes, as shown by their curves lying closer to the top-left corner. Ensemble techniques such as RF and XGBoost are also exhibited strong predictive outcomes, while k-NN and NB showed moderate discrimination ability.

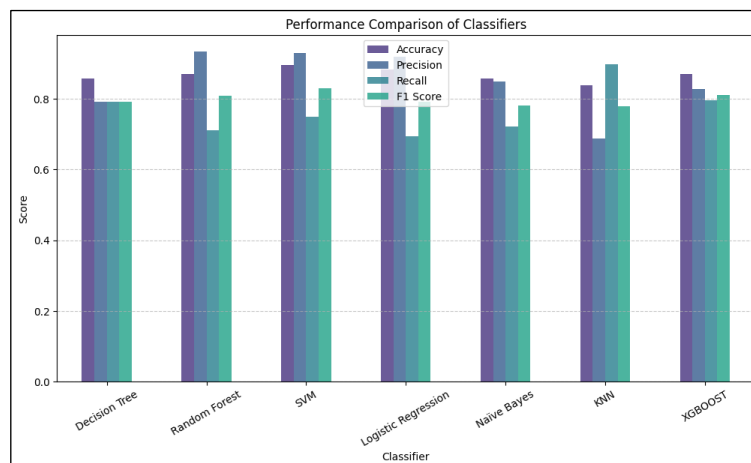


Figure 20: Performance statistics of seven classifiers.
Source: Authors, (2026).

Overall, as illustrated in Figure 17, SVM emerged as the most effective classifier for T2DM prediction. It is offered a good balance between sensitivity and specificity. LR and ensemble methods like RF and XGBoost also performed well, serving as strong alternatives. The k-NN algorithm showed high sensitivity but needs adjustment to lower false positives. Simpler models like DT and NB provided reasonable baseline results.

VII.1 CONFUSION MATRIX OF DIFFERENT CLASSIFIERS

Tables 4 to 10 present the confusion matrices for all seven classifiers applied to the PID dataset.

Table 4: DT Confusion Matrix.

Actual	Predicted		
	0	1	Σ
0	90 (89.10%)	11 (20.75%)	101
1	11 (10.89%)	42 (79.24%)	53
Σ	101	53	154

Source: Authors, (2026).

Table 5: RF Confusion Matrix.

Actual	Predicted		
	0	1	Σ
0	92 (84.40%)	3 (6.66%)	95
1	17 (15.59%)	42 (93.33%)	59
Σ	109	45	154

Source: Authors, (2026).

Table 6: SVM Confusion Matrix.

Actual	Predicted		
	0	1	Σ
0	99 (88.39%)	3 (7.14%)	102
1	13 (11.60%)	39 (92.85%)	52
Σ	112	42	154

Source: Authors, (2026).

Table 7: LR Confusion Matrix.

Actual	Predicted		
	0	1	Σ
0	102 (87.17%)	3 (8.10%)	105
1	15 (12.82%)	34 (91.89%)	49
Σ	117	37	154

Source: Authors, (2026).

Table 8: NB Confusion Matrix.

Actual	Predicted		
	0	1	Σ
0	93 (86.11%)	7 (15.21%)	100
1	15 (13.88%)	39 (84.47%)	54
Σ	108	46	154

Source: Authors, (2026).

Table 9: KNN Confusion Matrix.

Actual	Predicted		
	0	1	Σ
0	85 (94.44%)	20 (20.75%)	105
1	5 (5.55%)	44 (68.75%)	49
Σ	90	64	154

Source: Authors, (2026).

Table 10: XGBOOST Confusion Matrix.

Actual	Predicted		
	0	1	Σ
0	91 (89.21%)	9 (17.30%)	100
1	11 (10.78%)	43 (82.69%)	54
Σ	102	52	154

Source: Authors, (2026).

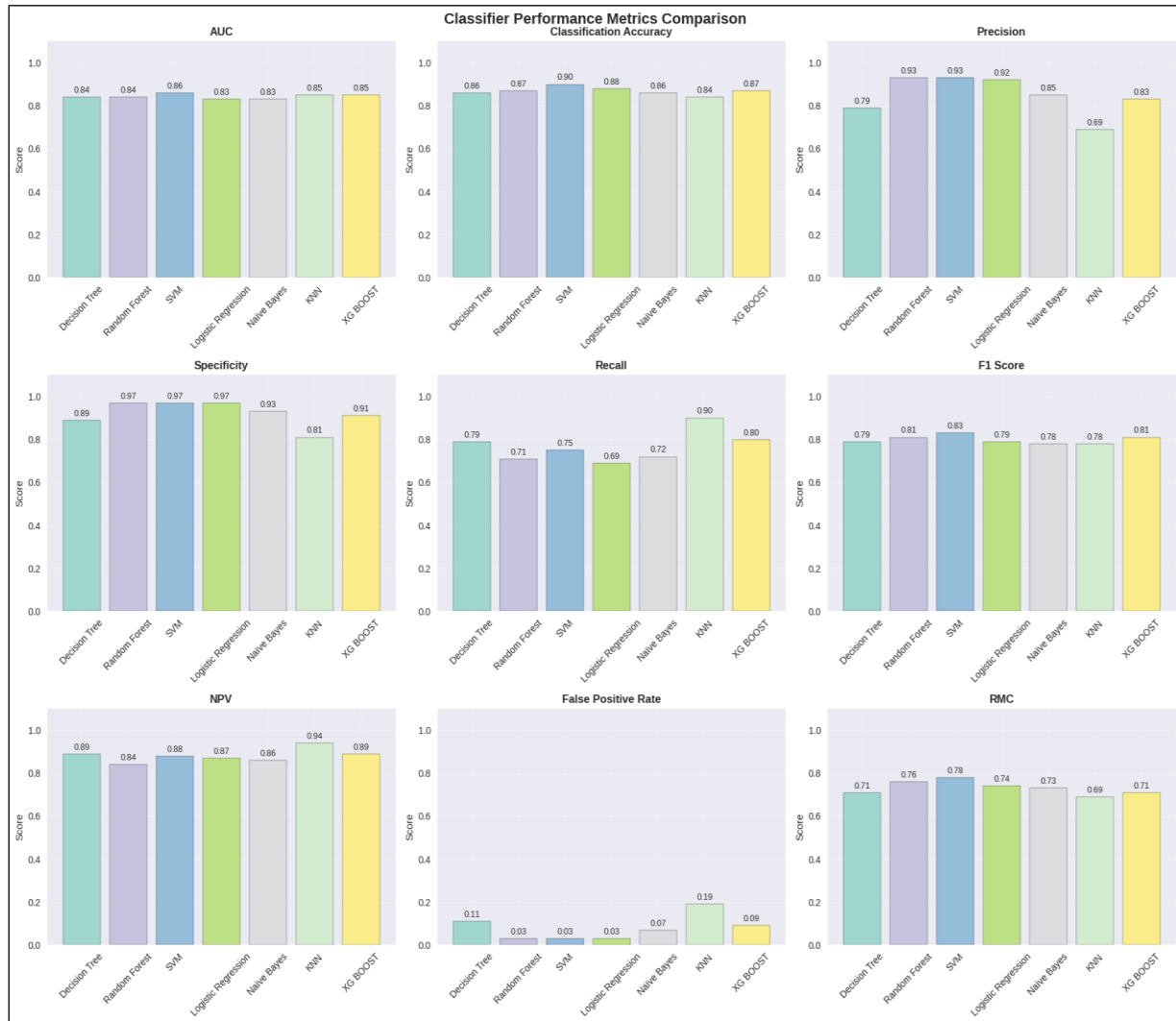


Figure 21: Classifiers performance metrics comparison.

Source: Authors, (2026).

Each table enumerates the instances classified as TP, FP, FN, and TN, presenting the counts alongside their corresponding percentages. A consolidated visual summary of the confusion matrices for all seven classifiers from Table 11 is provided in Figure 21.

Table 11: Summary of Classifier Performance Metrics.

S. N.	Classifier	AUC	CA	Precision	Specificity	Recall	F1	NPV	FPR	RMC
1	DT	0.84	0.86	0.79	0.89	0.79	0.79	0.89	0.11	0.71
2	RF	0.84	0.87	0.93	0.97	0.71	0.81	0.84	0.03	0.76
3	SVM	0.86	0.90	0.93	0.97	0.75	0.83	0.88	0.03	0.78
4	LR	0.83	0.88	0.92	0.97	0.69	0.79	0.87	0.03	0.74
5	NB	0.83	0.86	0.85	0.93	0.72	0.78	0.86	0.07	0.73
6	KNN	0.85	0.84	0.69	0.81	0.90	0.78	0.94	0.19	0.69
7	XG BOOST	0.85	0.87	0.83	0.91	0.80	0.81	0.89	0.09	0.71

Source: Authors, (2026).

VIII. CONCLUSION

This study conducted a systematic comparison of the accuracy and effectiveness of traditional and hybrid models for T2DM classification. Among the evaluated techniques, SVM consistently achieved superior discriminative ability, followed closely by LR, RF and XGBoost. These models demonstrated the ability to distinguish between diabetic and non-diabetic cases, while the ensemble methods providing a balanced combination of accuracy and stability. Although k-NN exhibited the highest recall, its tendency to generate false positives makes it less suitable as a standalone model. However, it is useful in situations where sensitivity is more important than specificity. The study highlights that selecting the most appropriate model should be guided by the intended clinical use, whether the priority is to maximize early detection or to reduce false alarms. Future extensions of this research may include integrating larger and more diverse datasets, applying deep learning frameworks, and exploring hybrid optimization strategies to further improve predictive reliability. By strengthening computational approaches to early diagnosis, this study contributes to the development of practical decision-support systems for diabetes care.

IX. AUTHOR'S CONTRIBUTION

Conceptualization: Rizwan Akhtar, Muhammad Kalamuddin Ahamad.

Methodology: Rizwan Akhtar and Muhammad Kalamuddin Ahamad.

Investigation: Rizwan Akhtar and Muhammad Kalamuddin Ahamad.

Discussion of results: Rizwan Akhtar, Muhammad Kalamuddin Ahamad.

Writing – Original Draft: Rizwan Akhtar.

Writing – Review and Editing: Rizwan Akhtar and Muhammad Kalamuddin Ahamad.

Resources: Muhammad Kalamuddin Ahamad.

Supervision: Muhammad Kalamuddin Ahamad.

Approval of the final text: Rizwan Akhtar, Muhammad Kalamuddin Ahamad.

X. ACKNOWLEDGMENTS

Statement of Approval. This study has been approved by Department of Computer Application, Integral University, Lucknow, India with Reference No: **IU/R&D/2026-MCN0004567**.

AVAILABILITY OF DATA AND MATERIALS

The data will be available to the readers of this research paper upon authentic requirement and request.

XI. REFERENCES

- [1] Iftikhar, K., Javaid, N., Ahmed, I., & Alrajeh, N. (2025). A novel explainable deep learning framework for accurate diabetes mellitus prediction. *Applied Sciences*, 15(16), 9162. <https://doi.org/10.3390/app15169162>
- [2] Islam, K., Islam, R., Nguyen, I., Malik, H., Pirzadah, H., Shrestha, B., Lentz, I. B., Shekoohi, S., & Kaye, A. D. (2025). Diabetes mellitus and associated vascular disease: pathogenesis, complications, and evolving treatments. *Advances in Therapy*. <https://doi.org/10.1007/s12325-025-03185-9>
- [3] Ayoade, O. B., Shahrestani, S., & Ruan, C. (2025). Machine Learning and Deep Learning Approaches for Predicting Diabetes progression: A Comparative analysis. *Electronics*, 14(13), 2583. <https://doi.org/10.3390/electronics14132583>
- [4] Alkhalefah, S., AlTuraiki, I., & Altwaijry, N. (2025). Advancing Diabetic Foot Ulcer Care: AI and generative AI approaches for classification, prediction, segmentation, and detection. *Healthcare*, 13(6), 648. <https://doi.org/10.3390/healthcare13060648>
- [5] Dey, A. K., Walia, P., Somvanshi, G., Ali, A., Das, S., Paul, P., & Ghosh, M. (2025, January 10). AI-Driven Diabetic Retinopathy Screening: Multicentric Validation of AIDRSS in India. *arXiv.org*. <https://arxiv.org/abs/2501.05826>
- [6] Nghiem, N., Wilson, N., Krebs, J., & Tran, T. (2024). Predicting the risk of diabetes complications using machine learning and social administrative data in a country with ethnic inequities in health: Aotearoa New Zealand. *BMC Medical Informatics and Decision Making*, 24(1). <https://doi.org/10.1186/s12911-024-02678-x>
- [7] American Diabetes Association. (2024). Standards of medical care in diabetes—2024. *Diabetes Care*, 47(Supplement_1), S1–S350. <https://doi.org/10.2337/dc24-S001>
- [8] S, G., Reddy, R. V. S., & Ahmed, M. R. (2023). Exploring the effectiveness of machine learning algorithms for early detection of Type-2 Diabetes Mellitus. *Measurement Sensors*, 31, 100983. <https://doi.org/10.1016/j.measen.2023.100983>
- [9] Yang, W., Wu, Y., Chen, Y., Chen, S., Gao, X., Wu, S., & Sun, L. (2024). Different levels of physical activity and risk of developing type 2 diabetes among adults with prediabetes: a population-based cohort study. *Nutrition Journal*, 23(1). <https://doi.org/10.1186/s12937-024-01013-4>
- [10] Yadu, S., Chandra, R., & Sinha, V. K. (2024). Comparing different machine learning techniques in predicting diabetes on early stage. *Engineering Proceedings*, 62, 20. <https://doi.org/10.3390/engproc2024062020>
- [11] Al-Dabbas, L., & Abu-Shareha, A. A. (2024). Early detection of female type-2 diabetes using machine learning and oversampling techniques. *Journal of Applied Data Sciences*, 5(3), 1237–1245. <https://doi.org/10.47738/jads.v5i3.298>
- [12] Ordoñez-Guillén, N. E., Gonzalez-Compean, J. L., Lopez-Arevalo, I., Contreras-Murillo, M., & Aldana-Bobadilla, E. (2023). Machine learning based study for the classification of Type 2 diabetes mellitus subtypes. *BioData Mining*, 16(1), 24. <https://doi.org/10.1186/s13040-023-00340-2>
- [13] Taha, A. A., & Malebary, S. J. (2022). A hybrid Meta-Classifer of fuzzy clustering and logistic regression for diabetes prediction. *Computers, Materials & Continua/Computers, Materials & Continua (Print)*, 71(3), 6089–6105. <https://doi.org/10.32604/cmc.2022.023848>
- [14] Wu, H., Yang, S., Huang, Z., He, J., & Wang, X. (2017). Type 2 diabetes mellitus prediction model based on data mining. *Informatics in Medicine Unlocked*, 10, 100–107. <https://doi.org/10.1016/j.imu.2017.12.006>
- [15] Alnowaiser, K. (2024). Improving healthcare prediction of diabetic patients using KNN imputed features and tri-ensemble model. *IEEE Access*, 12, 16783–16793. <https://doi.org/10.1109/ACCESS.2024.3359760>
- [16] Abdelhafez, H. A., & Amer, A. A. (2024). Machine learning techniques for diabetes prediction: A comparative analysis. *Journal of Applied Data Sciences*, 5(2), 792–807. <https://doi.org/10.47738/jads.v5i2.219>
- [17] Patro, K. K., Allam, J. P., Sanapala, U., Marpu, C. K., Samee, N. A., Alabdulhafith, M., & Plawiak, P. (2023). An effective correlation-based data modeling framework for automatic diabetes prediction using machine and deep learning techniques. *BMC Bioinformatics*, 24(1), 372. <https://doi.org/10.1186/s12859-023-05488-6>
- [18] Shaikh Bhat, S., Banu, M., Ansari, G. A., & Selvam, V. (2023). A risk assessment and prediction framework for diabetes mellitus using machine learning algorithms. *Healthcare Analytics*, 4(3), 100273. <https://doi.org/10.1016/j.health.2023.100273>
- [19] Afolabi, S., et al. (2025). Predicting diabetes using supervised machine learning. *Journal of Medical Systems*, 49(1), 13. <https://doi.org/10.1007/s10916-025-01772-2>

- [20] Kim, J., et al. (2024). Deep learning-based long-term risk evaluation of incident diabetes. *The Lancet Digital Health*, 6(3), e123-e131. [https://doi.org/10.1016/S2589-5370\(24\)00024-5](https://doi.org/10.1016/S2589-5370(24)00024-5)
- [21] Mohsen, F., et al. (2025). ECG features improve multimodal deep learning prediction of diabetes. *Scientific Reports*, 15(1), 12345. <https://doi.org/10.1038/s41598-025-12633-z>
- [22] Kulkarni, A. R., et al. (2023). Machine learning algorithm to non-invasively detect pre-diabetes using ECG data. *BMJ Innovations*, 9(1), 32-38. <https://doi.org/10.1136/bmjinnov-2022-000123>
- [23] Lee, H., et al. (2025). AI machine learning-based diabetes prediction in older adults. *JMIR Formative Research*, 9, e57874. <https://doi.org/10.2196/57874>
- [24] Abnoosian, K., et al. (2023). Prediction of diabetes disease using an ensemble machine learning approach. *Journal of Biomedical Informatics*, 130, 103984. <https://doi.org/10.1016/j.jbi.2022.103984>
- [25] Ansari, Y., et al. (2023). Deep learning for ECG arrhythmia detection and diabetes prediction. *Frontiers in Physiology*, 14, 1246746. <https://doi.org/10.3389/fphys.2023.1246746>
- [26] Alzboon, M. S., et al. (2025). A comparative study of machine learning techniques for early prediction of diabetes. *Journal of Healthcare Engineering*, 2025, 123456. <https://doi.org/10.1155/2025/123456>
- [27] Khokhar, P. B., et al. (2025). Towards transparent and accurate diabetes prediction using machine learning and explainable AI. *arXiv preprint arXiv:2501.18071*. <https://arxiv.org/abs/2501.18071>
- [28] Hennebelle, A., et al. (2023). HealthEdge: A machine learning-based smart healthcare framework for prediction of type 2 diabetes. *arXiv preprint arXiv:2301.10450*. <https://arxiv.org/abs/2301.10450>
- [29] Jena, S. K., Behera, D. K., Jena, A. K., & Rout, J. K. (2025). A novel ensemble machine learning framework for improved diabetes prediction and complication prevention. *Procedia Computer Science*, 258, 4008–4017. <https://doi.org/10.1016/j.procs.2025.04.652>
- [30] Alsadi, B., Musleh, S., Al-Absi, H. R. H., Refaee, M., Qureshi, R., El Hajj, N., & Alam, T. (2024). An ensemble-based machine learning model for predicting type 2 diabetes and its effect on bone health. *BMC Medical Informatics and Decision Making*, 24(1), 144. <https://doi.org/10.1186/s12911-024-02540-0>
- [31] Rastogi, R., & Bansal, M. (2022). Diabetes prediction model using data mining techniques. *Measurement Sensors*, 25, 100605. <https://doi.org/10.1016/j.measen.2022.100605>
- [32] Agliata, A., Giordano, D., Bardozzo, F., Bottiglieri, S., Facchiano, A., & Tagliaferri, R. (2023). Machine learning as a support for the diagnosis of Type 2 diabetes. *International Journal of Molecular Sciences*, 24(7), 6775. <https://doi.org/10.3390/ijms24076775>
- [33] Roobini, M. S., Lakshmi, M., Rajalakshmi, R., Sujihelen, L., & Babu, K. (2023). Type 2 diabetes mellitus classification using predictive supervised learning model. *Soft Computing*. <https://doi.org/10.1007/s00500-023-08726-4>
- [34] Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. *ICT Express*, 7(4), 432–439. <https://doi.org/10.1016/j.icte.2021.02.004>
- [35] Joshi, R. D., & Dhakal, C. K. (2021). Predicting Type 2 diabetes using logistic regression and machine learning approaches. *International Journal of Environmental Research and Public Health*, 18(14), 7346. <https://doi.org/10.3390/ijerph18147346>
- [36] Tigga, N. P., & Garg, S. (2020). Prediction of Type 2 Diabetes using Machine Learning Classification Methods. *Procedia Computer Science*, 167, 706–716. <https://doi.org/10.1016/j.procs.2020.03.336>
- [37] Alam, T. M., Iqbal, M. A., Ali, Y., Wahab, A., Ijaz, S., Baig, T. I., Hussain, A., Malik, M. A., Raza, M. M., Ibrar, S., & Abbas, Z. (2019). A model for early prediction of diabetes. *Informatics in Medicine Unlocked*, 16, 100204. <https://doi.org/10.1016/j.imu.2019.100204>
- [38] Mirzajani, S. S., & Salimi, S. (2018). Prediction and diagnosis of diabetes by using data mining techniques. *Avicenna Journal of Medical Biochemistry*, 6(1), 3–7. <https://doi.org/10.15171/ajmb.2018.02>
- [39] Mohammadi-Pirouz, Z., Hajian-Tilaki, K., Haddad-Zavareh, M. S., Amoozadeh, A., & Bahrami, S. (2024). Development of decision tree classification algorithms in predicting mortality of COVID-19 patients. *International Journal of Emergency Medicine*, 17(1). <https://doi.org/10.1186/s12245-024-00681-7>
- [40] Li, X., Ding, F., Zhang, L., Zhao, S., Hu, Z., Ma, Z., Li, F., Zhang, Y., Zhao, Y., & Zhao, Y. (2025). Interpretable machine learning method to predict the risk of pre-diabetes using a national-wide cross-sectional data: evidence from CHNS. *BMC Public Health*, 25(1). <https://doi.org/10.1186/s12889-025-22419-7>
- [41] Altman, N., & Krzywinski, M. (2017). Ensemble methods: bagging and random forests. *Nature Methods*, 14(10), 933–934. <https://doi.org/10.1038/nmeth.4438>
- [42] Badi, H., Patriciu, A., & Moutaouakil, K. E. (2024). Fractional Intuitionistic Fuzzy Support Vector Machine: Diabetes tweet Classification. *Information*, 15(11), 737. <https://doi.org/10.3390/info15110737>
- [43] Hennebelle, A., Materwala, H., & Ismail, L. (2023). HealthEdge: a Machine Learning-Based smart healthcare framework for prediction of Type 2 diabetes in an integrated IoT, edge, and cloud computing system. *arXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.2301.10450>
- [44] Aguilar-Ruiz, J. S., Romero, C., & Cicconardi, A. (2024, November 2). XNB: Explainable Class-Specific NAIVE-Bayes Classifier. *arXiv.org*. <https://arxiv.org/abs/2411.01203>
- [45] Halder, R. K., Uddin, M. N., Uddin, M. A., Aryal, S., & Khraisat, A. (2024). Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications. *Journal of Big Data*, 11(1). <https://doi.org/10.1186/s40537-024-00973-y>
- [46] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan and A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," in *IEEE Access*, vol. 8, pp. 107562-107582, 2020, doi: 10.1109/ACCESS.2020.3001149.
- [47] Dinanthi, D. A., Ramadanti, E., Aditya, C. S. K., & Chandranegara, D. R. (2024). Diabetes Detection Using Extreme Gradient Boosting (XGBoost) with Hyperparameter Tuning. *Indonesian Journal of Electronics Electromedical Engineering and Medical Informatics*, 6(2). <https://doi.org/10.35882/qr3hw926>
- [48] <https://www.kaggle.com/uciml/pima-indians-diabetes-database>