IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 30, No. 12, December, 2025, pp. 3235-3241
Journal homepage: http://iieta.org/journals/isi

A Comparative Analysis of Six Machine Learning Classifiers for Early-Stage Diabetes Risk Prediction Using a Kaggle Dataset

Areej Mahmoud Asaad[1*], Mahmood Hameed Qahtan[2], Ahmed Kh. Younis[1]

[1] Department of Computer Techniques Engineering, Technical Engineering College for Computer and Artificial Intelligence, Mosul, Northern Technical University, Mosul 41001, Iraq
[2] Department of Artificial Intelligence Techniques Engineering, Technical Engineering College for Computer and Artificial Intelligence, Mosul, Northern Technical University, Mosul 41001, Iraq

Corresponding Author Email: areej_mahmoud@ntu.edu.iq

https://doi.org/10.18280/isi.301216

Received: 3 November 2025
Revised: 22 December 2025
Accepted: 26 December 2025
Available online: 31 December 2025

Keywords:
diabetes prediction, comparative analysis, machine learning, ensemble methods, classification, medical informatics

ABSTRACT

Diabetes is a common metabolic condition characterized by an elevated blood sugar level due to impaired insulin production or action. Adverse sequelae of diabetes may be kidney damage, neuropathy, cardiovascular disease, and eye problems. Diabetes is increasingly becoming a regular phenomenon across the globe, and so, averting its impact on individuals and the healthcare systems will be to carry out early diagnosis of the disease, proper curative therapy, and preventive strategies. A study comparing various machine learning (ML) classifiers, including K-Nearest Neighbors (KNN), random forests (RF), Logistic Regression (LR), Gradient Boosting (GB), XGBoost, and decision trees (DT), was conducted to estimate the likelihood of diabetes. The model is evaluated by calculating accuracy, precision, recall, F1-score, execution time, and confusion matrix analysis. With the highest F1-score (0.99), accuracy (0.99), and recall (0.99), the Random Forest classifier performed exceptionally well, exhibiting remarkable resilience and classification capability. The accuracy, recall, and F1-score of both GB and XGBoost were 0.97, 0.96, and 0.97, respectively; however, XGBoost's execution time was longer than GB's. The decision tree model outperformed the LR model, achieving an accuracy of 0.92, a recall of 0.96, and an F1-score of 0.94. The decision tree model had an accuracy of 0.95, a recall of 0.93, and an F1-score of 0.95. The KNN model's accuracy, recall, and F1-score were 0.90, 0.89, and 0.93, respectively. With both high prediction accuracy and high sensitivity to positive cases, Random Forest is the best model for predicting diabetes overall, according to the data. This study makes it a good choice for applications needing early detection.

1. INTRODUCTION

Diabetes is a chronic insulin-related illness caused by either impaired insulin signalling or insufficient pancreatic insulin secretion [1]. This results in either low insulin production or inadequate insulin use by the body, leading to the accumulation of blood glucose, a characteristic of diabetes [2]. The World Health Organisation claims that diabetes is among the leading causes of death worldwide, and its manifestation is expected to rise in magnitude in the coming decades. With timely treatment, severe sequelae (cardiovascular diseases, kidney failure, neuropathy, etc.) can be prevented, which is also possible in time with early detection and a specific prediction [3, 4]. Advanced machine learning (ML) technology has been effective across various fields, including industry, education, and healthcare. An intelligent machine can imitate human behavior and is part of the field of artificial intelligence (AI).

AI systems can perform complex tasks, such as solving human problems. Managing different kinds using predictive analytics is one of the biggest applications of ML. Every instance in every dataset that ML algorithms employ is represented by the same collection of features, which can be categorical or continuous [5, 6]. Numerous ML classifiers, including Logistic Regression (LR), decision trees (DT), and random forests (RF), have been effectively used to forecast diabetes using patient data [7]. Using characteristics including age, body mass index (BMI), blood pressure, and glucose levels, these models categorise people as having diabetes or not [8]. The goal of this work is to provide a thorough comparative analysis of some of the most well-known ML classifiers for diabetes prediction, as well as the effectiveness of techniques including DT, RF, and LR. The study examines how feature selection strategies and data preprocessing affect the model's performance. The findings should enable the development of a robust disease-prediction model and provide insights into the advantages and disadvantages of various classifiers.

2. RELATED WORK

The early identification and management of diabetes on a global scale have unveiled profound opportunities in

healthcare informatics, particularly through the implementation of ML technologies [9]. This potential has prompted some scholars to investigate various ML classifiers to improve the precision and reliability of diabetes prediction systems. In 2023, Kangra and Singh [10] aimed to assess several ML techniques to achieve accurate diabetes forecasting. The six well-known classifiers used in their research were Support Vector Machine (SVM), Naïve Bayes (NB), K-Nearest Neighbors (KNN), RF, LR, and DT. In the WEKA 3.8.6 environment, experiments were carried out on two datasets: the German Diabetes Dataset (GDD) and the Pima Indian Diabetes Dataset (PIDD). They reported that KNN and RF outperformed other classifiers on the Germany dataset, achieving 98.7% accuracy, while SVM performed best overall on the PIDD, achieving 74% accuracy. The research showed that algorithm effectiveness varies across datasets for diabetes prediction, revealing algorithm performance metrics alongside error rates from competing classifiers. In 2024, Cichosz et al. [11] used uncontrolled diabetic data from the National Health and Nutrition Examination Survey (NHANES) from 2005 to 2018 to evaluate the predictive utility of five ML models for undiagnosed diabetes. Using biochemical verification on HbA1c levels, the study identified 45,431 individuals with previously undiagnosed diabetes from a large, diverse dataset. To assess the potential applicability of ML methods for prescreening, the authors focused on simple, readily available clinical variables. The examined models that included a neural network combined with Random Forest, AdaBoost, RUSBoost, and LogitBoost yielded AUCs between 0.776 and 0.806. Additionally, the models achieved high sensitivity rates of 0.742-0.871, alongside NPVs of 0.984-0.990.99. Though the positive predictive values. In 2024, Haji [12] attempted to establish a new model for predicting diabetes risk factors using SVMs on a publicly available Kaggle diabetes dataset. Along with other health indicators, the dataset included age, body mass index (BMI), and blood sugar levels. This study used an SVM classifier, feature selection, and intensive data preparation. Both training and validation were performed using traditional cross-validation to assess the model's reliability across datasets, a crucial step for models of this nature. The clinical outcomes were measured using F1-score, recall, precision, and accuracy.The test data accuracy of 83.12% confirms that the SVM model is a promising candidate for predicting diabetes risk from readily available clinical features. In 2025, Krishandhie and Purwinarko [13] conducted a study using the PIDD, preprocessing the data with SMOTE to balance the classes and using mean imputation for missing values (increasing the minority class from 268 to 454). The KNN and Random Forest algorithms are optimised. At 70:30, 75:25, and 80:20 ratios, the data used in this article is divided into training and test sets. Using a stacking ensemble strategy that combines KNN as the base classifier and RF as the meta-classifier to construct an RFKNN model, evaluated using confusion matrix analysis, yields the best accuracy of 92.86% on an 80:20 split. In 2024, Santiyuda [14] classified diabetes risk categories using the PIDD, implemented the KNN algorithm, and focused on data preparation, including preprocessing. To improve the input data for distance-based computation using the KNN method, missing-value imputation, normalization, and feature engineering were performed. To further enhance performance, some distance metrics, including Manhattan and Euclidean, were tested alongside hyperparameter tuning. Based on this study's findings, AI Detect Diabetes's general capabilities are limited, as evidenced by moderate accuracy (66%), precision (52%), and recall (58%), suggesting its inability to handle unbalanced datasets. Glucose levels and BMI were the most crucial characteristics. The research made clear that balanced datasets and more advanced feature selection methods are required. In 2025, Maulana et al. [15] examined an approach that uses homogeneous and heterogeneous methods. They used a dataset that included blood glucose and HbA1c values, age, gender, BMI, history of heart disease and hypertension, and smoking status. With balanced precision and recall, the best Boosted Random Forest model achieved 98% accuracy using AdaBoost and Random Forest as base estimators. Although it is marginally less accurate than the boosting strategy mentioned above, RF is also used as the base estimator in the bagging approach, which achieved 97% accuracy. The stacking approach achieves performance similar to boosted models while reducing prediction error, demonstrating its efficiency in terms of memory usage. It achieves comparable 98% accuracy but takes far less processing time, resulting in greater overall efficiency. In 2025, Zhu et al. [16] carried out a comparison of ensemble approaches, testing them on a 520-sample dataset with 17 features from the UCI ML Repository. Some features are basic behaviors, like age, gender, the index of obesity, smoking, and even drinking alcohol. In this work, SVM 1 is one of the three models utilized along with DT and LR. As indicated above, different pre-treatments (standardization and normalization) were applied to assess the model performance over various data representations. From the results of this study, SVM performed best among the other algorithms when trained on normalized datasets, achieving the highest recall, AUC, precision, accuracy, and F1-score. The raw dataset yielded inconsistent results across models: DT performed consistently, whereas LR showed only a single AUC peak.

In 2023, Al-Mousa et al. [17] outlined a diabetes detection technique that uses the CDC's Diabetes Health Indicators Dataset to categorise people as non-diabetic, pre-diabetic, or diabetic. 70% of the balanced dataset was used for training DT, K-Nearest Neighbours, RF, LR, and Stochastic Gradient Descent, while 30% was utilised for testing. Using 10-fold stratified cross-validation, performance was verified at 89% recall, accuracy, precision, and F1-score. The Random Forest classifier with 500 estimators outperformed decision tree (84%), KNN (82%), LR (58%), and SGD (54%). The model was robust, achieving 98% recall in classifying pre-diabetic cases. In 2025, Zhao [18] compared the K-Mean Clustering Algorithm and Random Forest Classifier Models holistically for diabetes Prediction using the PIDD. In addition to highlighting the clustering-based process and the precedence of an ideal ensemble method, this paper discusses the importance of early and accurate Detection of Diabetes. The K-means clustering algorithm achieved notable success, achieving 90.04% accuracy by splitting the data into meaningful parts based on intrinsic characteristics. However, Random Forest outperformed not only K-Means but also many other popular classifiers, including: KNN, Gradient Boosting (GB), DT, SVM, and even LR. Their findings underscored the usefulness of Random Forest for prediction and its potential practical applications in medicine. In 2025, Jena et al. [19] employed ML techniques using the 9-attribute PID dataset consisting of 768 cases. The data were collected from the UCI ML Repository, preprocessed using SMOTE to address class imbalance, and missing values were imputed using KNN. The

most relevant characteristics were chosen using Recursive Feature Elimination (RFE). Six ML approaches: a voting classifier ensemble that included LR, GB, SVM, RF, and DT, as well as another ensemble model with just RF and DT. The amalgamated ensemble demonstrated remarkable performance, achieving 84.2% accuracy. Although there are still certain limits, prior research indicates that ML can forecast diabetes. Numerous studies merely review a few algorithms without thoroughly evaluating their accuracy, computational efficiency, and execution time. Ensemble methods and rigorous validation approaches are understudied mainly, and hyperparameter optimisation is often overlooked. By contrasting six ML classifiers, this work seeks to address these issues. It uses many performance indicators, GridSearchCV optimisation, and standardised preprocessing. The findings provide valuable data for diabetes prediction in resource-constrained clinical settings. The following portions of this work are organised as follows: The technique, including data collection, preprocessing, and the ML methods used, is covered in Section 2. Section 3 displays the experiment's analysis and findings. Lastly, the key findings are summarized in the conclusion of Section 4.

3. METHODOLOGY

This section describes the methodology used to compare different ML classifiers for diabetes prediction. Data collection, preprocessing, model selection, hyperparameter tuning, and performance assessment are all included.

3.1 Datasets preprocessing

The system aims to estimate an individual's risk of developing diabetes using various demographic and clinical factors. For this analysis, the Early Stage Diabetes Risk Prediction Dataset available on Kaggle was selected [20]. This dataset is of moderate size, with 17 feature classes and 520 entries, which is suitable for ML applications. Genital thrush, partial paresis (muscle weakness), itching, irritability, delayed healing, muscle stiffness, alopecia (hair loss), age, gender, polyuria (excessive urination), polydipsia (excessive thirst), abrupt weight loss, weakness, and polyphagia (excessive hunger) are a few of these symptoms. The final property is the class label, which indicates if the individual has diabetes.

3.2 Algorithmic design of the ML model

The block diagram for the proposed system architecture is shown in Figure 1. The system uses multiple ML classifiers to predict diabetes in a systematic disciplined manner. Before the dataset was imported and preprocessed, gender and any symptom-related features (such as weakness, sudden weight loss, polyuria, and polydipsia, among others) were transformed into binary numerical values (Yes/Male/Positive = 1, No/Female/Negative = 0). The target variable (class) was encoded similarly, and any inaccurate or non-numeric inputs were considered missing values. Rows with missing values were removed to preserve data integrity. After the data was split into training and test sets (80/20), StandardScaler was used to standardise the features, remove the mean, and scale to unit variance. GridSearchCV was used with the KNN classifier due to its hyperparameter sensitivity. The remaining models were evaluated simultaneously with default parameters to maintain computational efficiency and provide a fair baselinecomparison.

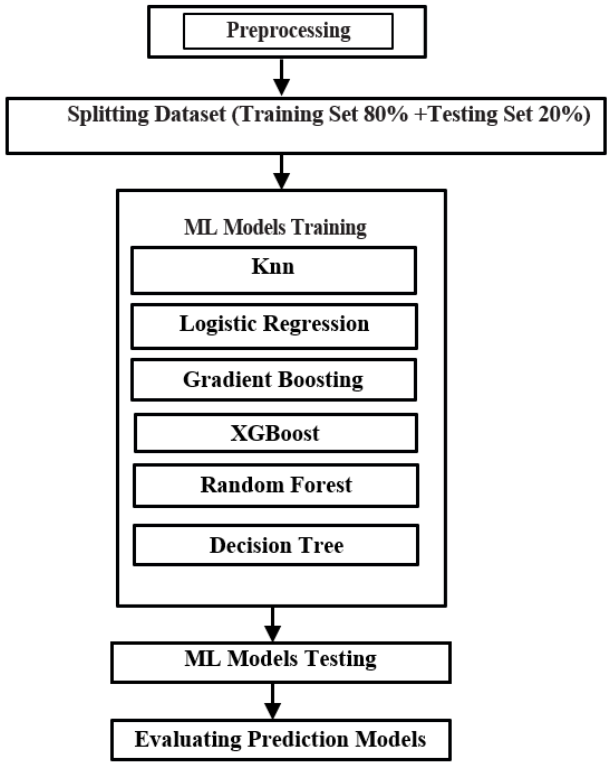

Figure 1. Block diagram of a machine learning (ML)-based diabetes prediction system

3.3 ML algorithms-based classification

3.3.1 Logistic Regression

A type of supervised ML model known primarily for handling binary classification challenges is LR [21]. Regression estimates the likelihood of an event and produces a 0 or 1 [22].

3.3.2 Decision tree

DT is a nonparametric model within the supervised learning framework and illustrates the metrics used to represent probabilistically different outcomes from training data inputs using a tree-like diagram [23].

3.3.3 Random forests

One of the most well-known approaches to ML in data mining is the RF method. Random Forest uses an ensemble approach and has gained significant notoriety for its extreme usefulness [24].

3.3.4 K-Nearest Neighbors

An algorithm works on very straightforward ideas, such as classifying new information using labeled training information anticipating that coinciding information from a specific record will be found within it, and predicting its category, KNN was used in this case, so it is set to a distance from various points relative to the other records [25, 26].

3.3.5 XGBoost and Gradient Boosting

Extreme GB is a fascinating blend of gradient descent and boosting techniques, often referred to as the Gradient Boosting Machine (GBM). Boosting is an ensemble learning method

that updates the weights of the training data at each learning round. In each boosting round, it increases the weights of misclassified samples and decreases those of correctly classified samples, effectively altering the training data distribution [27, 28].

4. RESULTS AND ANALYSIS

4.1 Evaluation of model performance

A brief overview of the key performance measures for evaluating the model, accuracy, recall, precision, F1-score, and Confusion Matrix (CM), is provided in this section. These measures are crucial for assessing the effectiveness of a system to classify illnesses. Each of these measures uses values for true positives (TP), false positives (FP), false negatives (FP), and true negatives (TN) [29].

The confusion matrix visually illustrates where errors occur in predictions, providing a clear picture of a model's performance. The predicted class labels appear in the columns, and the actual class labels appear in the rows, providing information on the types and frequencies of misclassifications [30].

Accuracy is the most common measure for evaluating a system's performance. In essence, it is the proportion of the correctly predicted instances to the total number of predictions, as expressed in the equation below [31]:

$$\text{Accuracy} = \text{TP} + \text{TN}/\text{TP} + \text{FP} + \text{TN} + \text{FN} \quad (1)$$

Recall is a measure of how well the model detects actual positive cases, or, put more simply, how many of the actual positive cases the model detects. Recall is calculated using the formula provided below [32].

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

Precision: the proportion of true positives among all positives. It is the number of accurate optimistic predictions divided by the number of instances predicted as positive, as shown in the formula below [33]:

$$\text{Precision} = \text{TP}/\text{TP} + \text{FP} \quad (3)$$

F1-score: is a satisfactory score that takes into consideration both precision and recall, and provides one single number to represent the overall performance score. It can be helpful when an even balance between precision and recall is necessary. The formula is shown below [34-36]:

$$\text{F1-score} = \text{Precision}*\text{Recall} / \text{Precision} + \text{Recall} \quad (4)$$

Random Forest achieved the highest numerical performance among all evaluated models. However, McNemar's test showed that it was statistically superior only to LR ($p = 0.0391$). In contrast, no statistically significant differences were observed between Random Forest and other tree-based or ensemble models, including GB ($p = 0.5000$), XGBoost ($p = 0.6250$), and Decision Tree ($p = 0.2188$). These findings indicate that GB, XGBoost, and Decision Tree exhibit performance levels that are closely comparable to those of Random Forest. Overall, the results demonstrate that ensemble and tree-based models are highly effective for diabetes prediction on this dataset.

4.2 K-Nearest Neighbors

Figure 2 shows that the KNN model has balanced performance but can be sensitive to data scaling and the choice of k. Strengths: Performs well with optimized k (via GridSearchCV). Matrix values: Top-left (TN): 31 correctly classified negative instances. Top-right (FP): Two cases were mislabeled as positive when they were actually negative.Bottom-left (FN): Eight cases that were mistakenly categorised as negative but were really positive.Bottom-right (TP): 63 Correctly classified positive instances. The model achieves approximately 90.4% accuracy, indicating it correctly classifies the majority of cases.

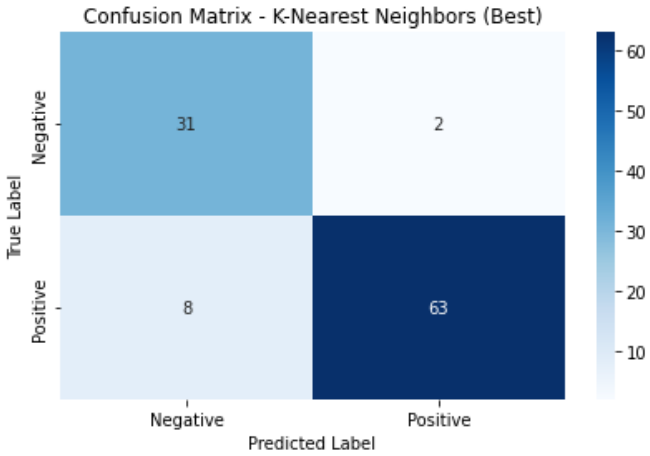

Figure 2. Confusion Matrix (CM) of K-Nearest Neighbors (KNN) model

4.3 Random forests

Figure 3 shows a Random Forest classifier used to evaluate performance. Matrix values: Top-left (TN): 33 correctly classified negative instances. Top-right (FP): 0 Instances incorrectly classified as positive when they are negative. Bottom-left (FN): 1 Instance incorrectly classified as negative when it is positive—Bottom-right (TP): 70 correctly classified positive instances. The model achieves approximately 99.0% accuracy, indicating extremely high overall performance.

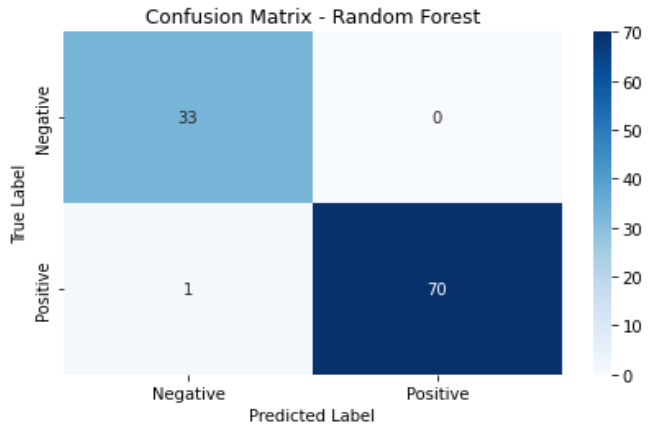

Figure 3. Confusion Matrix (CM) of the random forests (RF) model

4.4 Logistic Regression

Figure 4 shows an LR algorithm for evaluating performance. Matrix values: Top-left (TN): 28 correctly classified negative cases. Top-right (FP): 5 instances incorrectly classified as positive when they are negative. Bottom-left (FN): 3 instances incorrectly classified as negative when they are positive. Bottom-right (TP): 68 correctly classified positive instances. The LR model achieves an accuracy of 92.3%, indicating that it performs well overall in correctly classifying the cases.

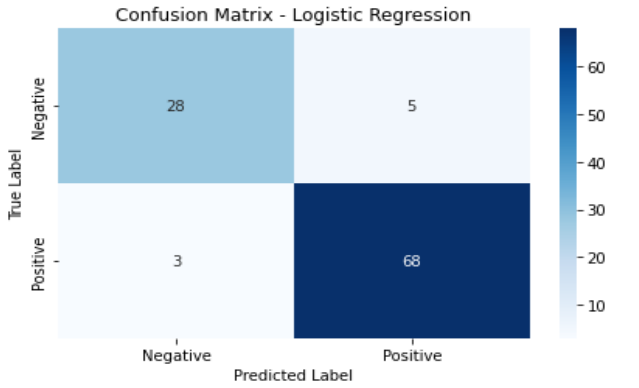

Figure 4. Confusion Matrix (CM) of Logistic Regression (LR) model

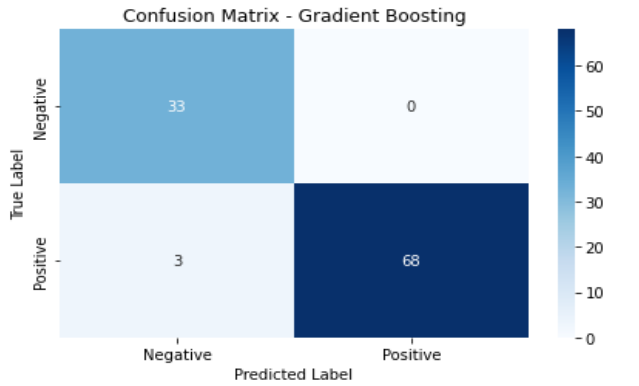

Figure 5. Confusion Matrix (CM) of the Gradient Boosting (GB) model

4.5 Gradient Boosting

Figure 5 illustrates a GB classifier to evaluate the performance. Matrix values: Top-left (TN): 33 Correctly classified negative instances.Top-right (FP): 0 Instances incorrectly classified as positive when they are negative.Bottom-left (FN): 3 Instances incorrectly classified as negative when they are positive.Bottom-right (TP): 68Correctly classified positive instances. The model achieves 97.1% accuracy, reflecting excellent overall classification performance.

4.6 XGBoost

Figure 6 shows the XGBoost model to evaluate its performance. Matrix values: Top-left (TN): 33 correctly classified negative instances. Top-right (FP): 0 instances incorrectly classified as positive when they are negative.Bottom-left (FN): 3 instances incorrectly classified as negative when they are positive.Bottom-right (TP): 68 correctly classified positive instances. The model achieves 97.1% accuracy, demonstrating strong overall performance.

4.7 Decision tree

Figure 7 shows a DT model for evaluating performance. Matrix values: Top-Left (TN): 33 correctly predicted negative. Top-Right (FP): 0 incorrectly predicted positive. Bottom-Left (FN): 5 incorrectly predicted negatives. Bottom-Right (TP): 66 correctly predicted positives. The DT algorithm achieves 95.2% accuracy, indicating solid, reliable classification performance.

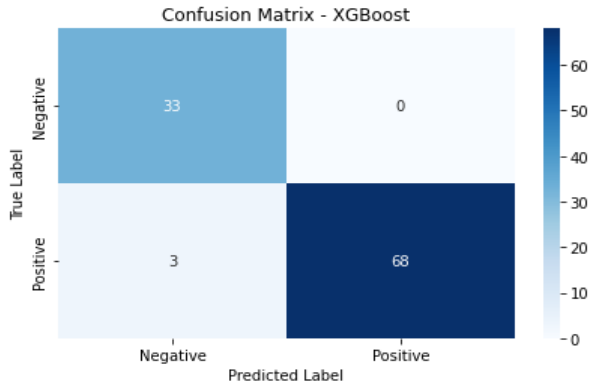

Figure 6. Confusion Matrix (CM) of XGBoost model

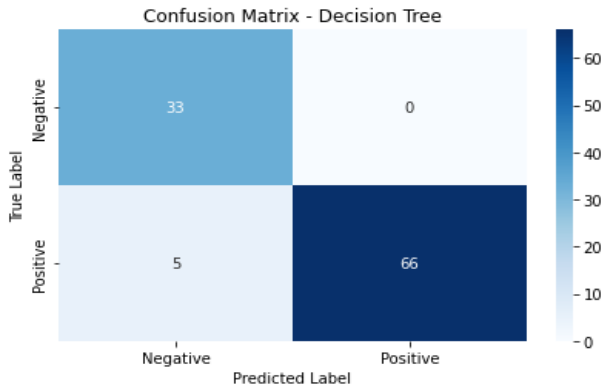

Figure 7. Confusion Matrix (CM) of the Decision Tree (DT) model

Table 1. Performance comparison of machine learning (ML) models for diabetes prediction

Model	Accuracy	Recall	Precision	F1-Score	Time (s)
KNN (Best)	0.90	0.89	0.91	0.91	0.04
Random Forest	0.99	0.99	0.99	0.99	0.17
Logistic Regression (LR)	0.92	0.96	0.92	0.92	0.03
Gradient Boosting	0.97	0.96	0.97	0.97	0.17
XGBoost	0.97	0.96	0.97	0.97	1.08
Decision Tree (DT)	0.95	0.93	0.96	0.95	0.00

Table 1 compares the performance of six different ML models across three key metrics: Accuracy, F1-score, and Execution Time (in seconds).

In Figure 8, the chart presents a comparison of ML models based on Accuracy, F1-score, and Execution Time, where:

- **A Bar Plot** illustrated the precise Accuracy and F1-score for each algorithm.
- **The Red Line Plot** represented the execution time in seconds.

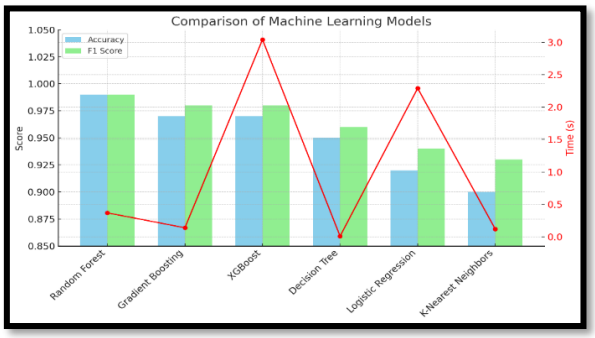

Figure 8. Evaluation of ML models: Accuracy, F1-score, and execution time

5. CONCLUSION

This model has proved efficient in ML for classifying the type of diabetes using medical and other lifestyle features. The Random Forest model was the best classifier among those tested, achieving an accuracy of 0.99, making it well-suited for application in clinical decision support systems. GB and XGBoost achieved high performance, but the latter was much more computationally intensive. 'Standard' models like LR and DT worked pretty well, whereas KNN was comparatively less accurate even after GridSearchCV optimization. Results highlight the promise of ensemble learning methods, particularly Random Forest, in medical prediction. Based on McNemar's statistical analysis, tree-based and ensemble classifiers provide the most reliable performance for diabetes prediction, outperforming LR while showing no significant differences among themselves. In the future, implementing such models in clinical practice could enhance sensitivity for early detection and improve the management of diabetes.

REFERENCES

[1] Petridis, P.D., Kristo, A.S., Sikalidis, A.K., Kitsas, I.K. (2024). A review on trending machine learning techniques for type 2 diabetes mellitus management. Informatics, 11(4): 70. https://doi.org/10.3390/informatics11040070

[2] Mishra, S., Tiwari, P., Yadav, R., Patel, P.S. (2024). An extensive analysis of diseases associated with diabetes. Journal of Pharma Insights and Research, 2(3): 174-187. https://doi.org/10.69613/ng1j7s13

[3] Abbasi, S., Mir, U.Y., Parveen, S.A., Azeez, A. (2024). A study on the prevalence of type 2 diabetes in OPD of national institute of Unani medicine Bengaluru, India. International Journal of Recent Scientific Research, 15(7): 1-4.

[4] Zhang, Z.W., Deng, C., Paulus, Y.M. (2024). Advances in structural and functional retinal imaging and biomarkers for early detection of diabetic retinopathy. Biomedicines, 12(7): 1405. https://doi.org/10.3390/biomedicines12071405

[5] Pallikonda, A.K., Bandarapalli, V.K., Aruna, V. (2025). Artificial intelligence and machine learning in smart healthcare: Advancing patient care and medical decision-making. Healthcraft Frontiers, 3(1): 47-57. https://doi.org/10.56578/hf030105

[6] Mustafa, M.A.S. (2025). Predictive reliability-driven optimization of spare parts management in aircraft fleets using AI, IoT, and digital twin technologies. Journal of Engineering Management and Systems Engineering, 4(3): 218-236. https://doi.org/10.56578/jemse040305

[7] Kim, B. (2024). A study on diabetes management system based on logistic regression and Random Forest. International Journal of Advanced Smart Convergence, 13(2): 61-68. http://doi.org/10.7236/IJASC.2024.13.2.61

[8] Jain, R., Tripathi, N.K., Pant, M., Anutariya, C., Silpasuwanchai, C. (2024). Investigating gender and age variability in diabetes prediction: A multi-model ensemble learning approach. IEEE Access, 12: 71535-71554. http://doi.org/10.1109/ACCESS.2024.3402350

[9] Al-Karakchi, A.A.A., Albanna, E., Al-Rifaie, A.H. (2023). Dynamic voltage restorer for voltage unbalance mitigation and voltage profile improvement in distribution network. Przegląd Elektrotechniczny, 99(6): 188-191. https://doi.org/10.15199/48.2023.06.38

[10] Kangra, K., Singh, J. (2023). Comparative analysis of predictive machine learning algorithms for diabetes mellitus. Bulletin of Electrical Engineering and Informatics, 12(3): 1728-1737. http://doi.org/10.11591/eei.v12i3.4412

[11] Cichosz, S.L., Bender, C., Hejlesen, O. (2024). A comparative analysis of machine learning models for the detection of undiagnosed diabetes patients. Diabetology, 5(1): 1-11. https://doi.org/10.3390/diabetology5010001

[12] Haji, V.M. (2024). Enhanced SVM classification for diabetes prediction: A comparative analysis using the kaggles diabetes dataset. Communications on Applied Nonlinear Analysis, 31(8s): 237-247. https://doi.org/10.52783/cana.v31.1477

[13] Krishandhie, S.Z.R., Purwinarko, A. (2025). Random Forest algorithm optimization using K-Nearest Neighbor and SMOTE on diabetes disease. Recursive Journal of Informatics, 3(1): 43-50. https://doi.org/10.15294/rji.v3i1.1576

[14] Santiyuda, K.G. (2024). K-Nearest Neighbors approach to classify diabetes risk categories. Jurnal Sistem Informasi dan Komputer terapan Indonesia (JSIKTI), 7(2): 74-83. https://doi.org/10.33173/jsikti.197

[15] Maulana, M.N., Muljono, Meindiawan, E.P.A. (2025). Comparative analysis of homogeneous and heterogeneous ensembles for diabetes classification optimization. Sinkron: Jurnal Dan Penelitian Teknik Informatika, 9(1): 512-521. https://doi.org/10.33395/sinkron.v9i1.14439

[16] Zhu, X.L., Liao, H.T., Zhang, T.T., Zhang, L.J., Li, X.R. (2025). A comparative study of machine learning in diabetes classification: Decision trees, logistic regression, and SVM. In Second International Conference on Big Data, Computational Intelligence, and Applications. http://doi.org/10.1117/12.3059546

[17] Al-Mousa, A., AlKhdour, L., Bishawi, H., AlShubeliat, F. (2023). Multiclass diabetes detection using Random Forest classification. In 2023 IEEE Jordan International Joint Conference on Electrical Engineering and

Information Technology (JEEIT), Amman, Jordan, pp. 243-248. https://doi.org/10.1109/JEEIT58638.2023.10185679
[18] Zhao, Y. (2025). Comparative analysis of diabetes prediction models using the Pima Indian Diabetes Database. ITM Web of Conferences, 70: 02021. https://doi.org/10.1051/itmconf/20257002021
[19] 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
[20] Kaggle: Early stage diabetes risk prediction dataset. https://www.kaggle.com/datasets/ishandutta/early-stage-diabetes-risk-prediction-dataset.
[21] Hasan, S.Q. (2023). Shallow model and deep learning model for features extraction of images. NTU Journal of Engineering and Technology, 2(3): 1-8. https://doi.org/10.56286/ntujet.v2i3.449
[22] Patil, M.S., Patil, H.D. (2024). Logistic regression based model for pain intensity level detection from biomedical signal. International Research Journal of Multidisciplinary Scope (IRJMS), 5(2): 652-662. https://doi.org/10.47857/irjms.2024.v05i02.0595
[23] Coscia, A., Dentamaro, V., Galantucci, S., Maci, A., Pirlo, G. (2024). Automatic decision tree-based NIDPS ruleset generation for DoS/DDoS attacks. Journal of Information Security and Applications, 82: 103736. https://doi.org/10.1016/j.jisa.2024.103736
[24] Salman, H.A., Kalakech, A., Steiti, A. (2024). Random Forest algorithm overview. Babylonian Journal of Machine Learning, 2024: 69-79. https://doi.org/10.58496/BJML/2024/007
[25] Younis, A.K., Younis, B.M., Jarjees, M.S. (2022). Hardware implementation of a Sobel edge detection system for a blood cell image-based field programmable gate array. Indonesian Journal of Electrical Engineering and Computer Science, 26(1): 86-95. https://doi.org/10.11591/ijeecs.v26.i1.pp86-95
[26] Moon, K., Jetawat, A. (2024). Predicting lung cancer with K-Nearest Neighbors (KNN): A computational approach. Indian Journal of Science and Technology, 17(21): 2199-2206. https://doi.org/10.17485/IJST/v17i21.1192
[27] Hanif, I. (2020). Implementing Extreme Gradient Boosting (XGBoost) classifier to improve customer churn prediction. In Proceedings of the 1st International Conference on Statistics and Analytics, Bogor, Indonesia, pp. 1-20. https://doi.org/10.4108/eai.2-8-2019.2290338
[28] Islam, S.F.N., Sholahuddin, A., Abdullah, A.S. (2021). Extreme Gradient Boosting (XGBoost) method in making forecasting application and analysis of USD exchange rates against rupiah. Journal of Physics: Conference Series, 1722: 012016. https://doi.org/10.1088/1742-6596/1722/1/012016
[29] Qahtan, M.H., Asaad, A.M., Younis, A.K. (2025). Bone fracture detection using hybrid EfficientNet-B0 and ResNet50 with SVM: A comparative performance analysis. Ingénierie des Systèmes d'Information, 30(7): 1775-1782. https://doi.org/10.18280/isi.300710
[30] Berliana, E.V., Riasetiawan, M. (2024). Comparative analysis of Naïve Bayes classifier, support vector machine, and decision tree in rainfall classification using confusion matrix. International Journal of Advanced Computer Science and Applications (ijacsa), 15(7): 560-567. https://doi.org/10.14569/IJACSA.2024.0150755
[31] Najam, N.R., Abduljawad, R.A. (2023). RF-RFE-SMOTE: A DoS and DDoS attack detection framework. NTU Journal of Engineering and Technology, 2(2): 29-47. https://doi.org/10.56286/ntujet.v2i2.436
[32] Sathyanarayanan, S., Tantri, B.R. (2024). Confusion matrix-based performance evaluation metrics. African Journal of Biomedical Research, 27(4s): 4023-4031. https://doi.org/10.53555/AJBR.v27i4S.4345
[33] Shirdel, M., Di Mauro, M., Liotta, A. (2024). Worthiness benchmark: A novel concept for analyzing binary classification evaluation metrics. Information Sciences, 678: 120882. https://doi.org/10.1016/j.ins.2024.120882
[34] El Emary, I.M.M., Yaghi, K.A. (2024). Machine learning classifier algorithms for ransomware lockbit prediction. Journal of Applied Data Sciences, 5(1): 24-32. https://doi.org/10.47738/jads.v5i1.161
[35] Alhelal, D., Younis, A.K., Al-Mallah, R.H.A. (2021). Detection of brain stroke in the MRI image using FPGA. TELKOMNIKA Telecommunication, Computing, Electronics and Control, 19(4): 1307-1315. https://doi.org/10.12928/TELKOMNIKA.v19i4.18988
[36] Ercan, T., Al Azzawi, A.K. (2019). Design of an FPGA-based intelligent gateway for industrial IoT. International Journal of Advanced Trends in Computer Science and Engineering, 8(1.2): 126-130. https://www.warse.org/IJATCSE/static/pdf/file/ijatcse21812sl2019.pdf.