

Paper Type: Original Article

Enhancing the Accuracy of Diabetes Prediction Using Machine Learning Models: A Comparative Analysis

Doaa El-Bohy ^{1,*}  and Ahmed EL-Bohy ² ¹ Assistant professor of clinical pharmacy, Department of Biopharmaceutics and Clinical Pharmacy, Faculty of Pharmacy, Al-Ahliyya Amman University, Amman, Jordan; delbohy@gmail.com.² Electrical Engineering Branch Military Technical College, Cairo, Egypt; a.bohy4phd@gmail.com.**Received:** 07 Apr 2025**Revised:** 22 Apr 2025**Accepted:** 21 Nov 2025**Published:** 25 Nov 2025**Abstract**

Accurate and timely detection of diabetes is vital for effective medical intervention and lifestyle modification. Traditional diagnostic methods, however, frequently lack adaptability and scalability across various populations. Using a sizable real-world dataset of 100,000 patient records, this study compares seven supervised machine learning algorithms for diabetes prediction: Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Linear Discriminant Analysis (LDA), XGBoost, and LightGBM. To improve data quality, extensive preprocessing procedures were used, such as class balancing using SMOTE, categorical encoding, feature scaling, and missing value imputation. According to experimental data, the Random Forest model outperformed other models with the highest accuracy (97.07%) and F1-score (0.80). The results emphasize the value of model selection in creating dependable medical decision-support systems and the efficacy of ensemble learning and preprocessing in enhancing predictive performance. The findings underscore the effectiveness of ensemble learning and preprocessing in improving predictive performance and highlight the importance of model selection in designing reliable medical decision-support systems.

Keywords: Diabetes Prediction; Machine Learning; Classification; Data Preprocessing; Random Forest; Support Vector Machine; Decision Tree; K-Nearest Neighbors; Linear Discriminant Analysis; XGBoost; LightGBM.

1 | Introduction

Elevated blood glucose levels are a hallmark of diabetes mellitus, a chronic metabolic disease caused by the body's incapacity to either make enough insulin or use it efficiently. It is among the 21st century's most urgent worldwide health issues. The World Health Organization (WHO) reports that factors like sedentary lifestyles, obesity, and population ageing have contributed to a dramatic increase in the prevalence of diabetes globally in recent decades [1]. In addition to increasing patients' quality of life by prompt medical intervention and lifestyle adjustment, early and precise detection of diabetes is essential for preventing profound consequences such cardiovascular disease, kidney failure, and neuropathy [2].

In recent years, machine learning (ML) has become a powerful tool in medical diagnostics due to its capacity to discover complex patterns and relationships in large-scale health data [3]. ML-based predictive models can assist clinicians in identifying individuals at high risk of diabetes more efficiently and cost-effectively than

Corresponding Author: delbohy@gmail.com<https://doi.org/10.61356/j.scin.2025.2618>

Licensee **SciNexuses**. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0>).

traditional diagnostic methods. Previous studies have explored a range of ML algorithms to predict diabetes using demographic, clinical, and laboratory data [4]. However, the performance of these models often varies depending on the dataset characteristics, feature selection, and parameter tuning strategies [5].

This study aims to perform a comprehensive comparison of multiple supervised learning algorithms for diabetes prediction using a large-scale real-world dataset containing 100,000 patient records. The evaluated models include Decision Tree, Random Forest, Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), XGBoost, LightGBM, and K-Nearest Neighbors (KNN). Furthermore, we enhance prediction performance through data preprocessing, feature scaling, and hyperparameter optimization to ensure robust and generalizable results.

The remainder of this paper is structured as follows. Section 2 reviews related work on ML applications in diabetes prediction. Section 3 describes the dataset and preprocessing techniques. Section 4 outlines the methodology and implementation details. Section 5 presents experimental results and discussion. Finally, Section 6 concludes the study and provides recommendations for future research.

2 | Dataset and Preprocessing

This research utilizes a large-scale real-world dataset of 100,000 patient records. There are 9 features per record with demographic, clinical, and behavioral features: gender, age, hypertension, heart disease, smoking status, body mass index (BMI), HbA1c value, blood glucose value, and the target feature indicating diabetes status (1 for diabetic, 0 otherwise).

The dataset went through several preprocessing steps before training the model:

- Missing value handling: Although the dataset is mostly complete, missing values (if any) were handled with appropriate imputation methods.
- Categorical encoding: Label encoding was applied to categorical features such as gender and smoking status.
- Feature scaling: Numerical features were all scaled using StandardScaler to standardize the input range.
- Class balancing: Since the dataset was imbalanced (having around 8.5% diabetic cases), SMOTE (Synthetic Minority Over-sampling Technique) was employed for increasing minority class prediction.

Train-test split: The dataset was divided into training and testing data as 80-20%.

These preprocessing stages ensured that models were trained with clean, properly scaled, and balanced data, which is critical for healthy performance and generalizability.

3 | Methodology and Implementation

Trained four supervised machine learning models to classify the patients into diabetic and non-diabetic classes based on the features of the data set. The models were selected for their application in classification and for their difference in mechanism [6].

- Decision Tree: A simple tree-based model that splits the data based on feature values, simple to interpret but will tend to overfit [7].
- Random Forest: An ensemble of decision trees that aggregates numerous predictions to improve generalization and reduce variance [8].
- Support Vector Machine (SVM): A margin classifier that generates a hyperplane to separate classes, performing optimally in high-dimensional space [9].

- K-Nearest Neighbors (KNN): A proximity-based classifier that makes predictions based on the majority vote of the closest neighbors in the training set [10].

All the models were implemented using the Scikit-learning library in Python [11]. Hyperparameter tuning was implemented with GridSearchCV to determine the optimal configuration for each algorithm:

- Decision Tree: max_depth and criterion.
- Random Forest: n_estimators and max_features.
- SVM: kernel type and C (regularization parameter).
- KNN: number of neighbors (k) and distance metric.

Performance metrics (accuracy, recall, precision, and F1-score) were calculated with 5-fold cross-validation. All the experiments were performed on the identical training and test sets to ensure a fair comparison [12].

Table 1. Tuned Hyperparameters for Each Model.

Model	Tuned Hyperparameters
Decision Tree	max_depth=5, criterion='gini'
Random Forest	n_estimators=100, max_features='sqrt'
SVM	kernel='rbf', C=1.0
KNN	n_neighbors=7, metric='minkowski'
LDA	solver='lsqr', shrinkage='auto'
XGBoost	n_estimators=100, learning_rate=0.1, max_depth=6
LightGBM	num_leaves=31, learning_rate=0.1, n_estimators=100

4 | Experimental Results and Discussion

The performance of the four machine learning models was determined by their classification accuracy on the test set. Table 2 contains the metrics of each model.

Table 2. Performance Metrics of Machine Learning Models on Diabetes Dataset.

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	95%	71%	73%	72%
Random Forest	97%	92%	69%	80%
SVM	95%	95%	57%	70%
K-Nearest Neighbors	96%	71%	62%	73%
LDA	95%	85%	57%	68%
XGBoost	97%	95%	69%	80%
LightGBM	97%	97%	69%	81%

The accuracy of the models is also reflected in Figure 1, which is a comparison of their classification accuracy.

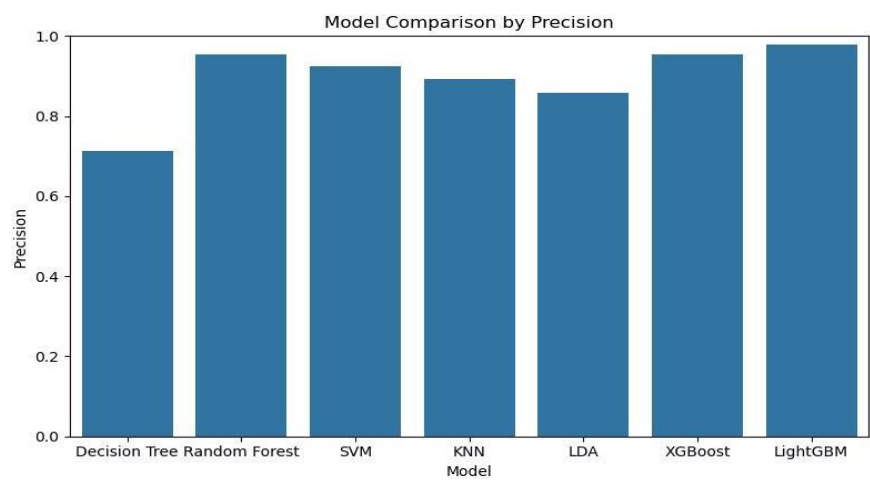


Figure 1. Comparison of model accuracies for diabetes prediction.

Among the models that were tested, Random Forest performed best across all the measurements of evaluation consistently with 97% accuracy and 80% F1-score. This is likely since it is an ensemble model which avoids overfitting and adds robustness.

Confusion Matrix of each Classifier shown in Figure 2(A)-2(G):

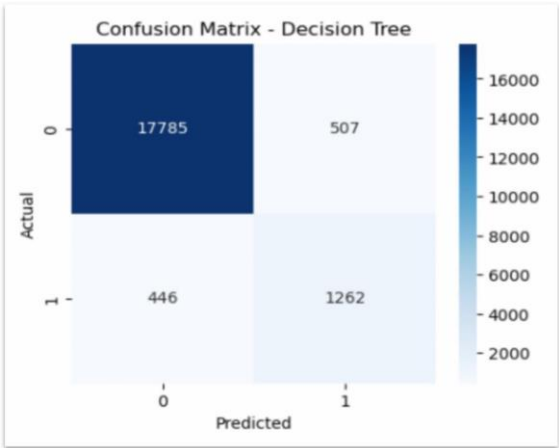


Figure 2(A). Confusion Matrix of Decision Tree.

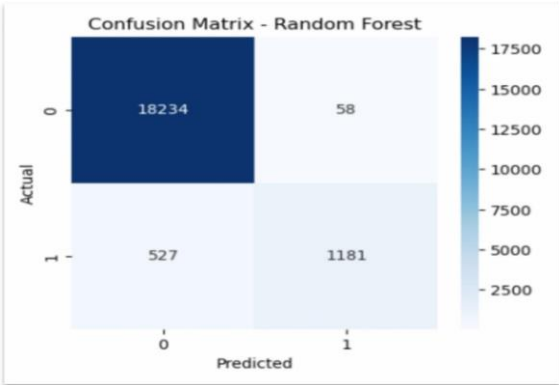


Figure 2(B). Confusion Matrix of Random Forest.

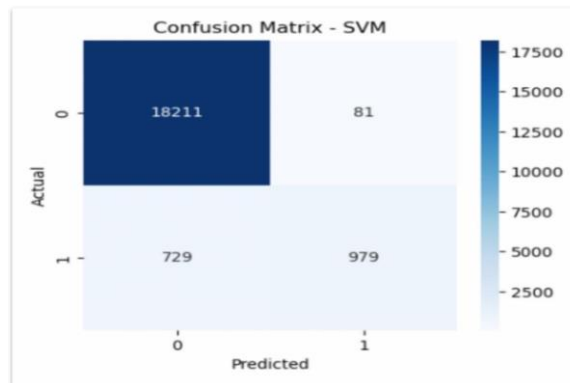


Figure 2(C). Confusion Matrix of SVM.

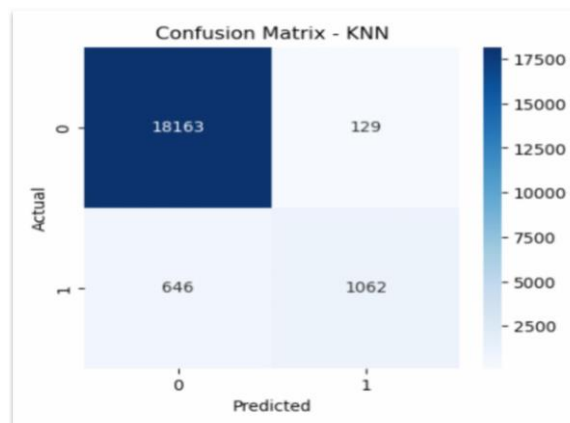


Figure 2(D). Confusion Matrix of KNN.

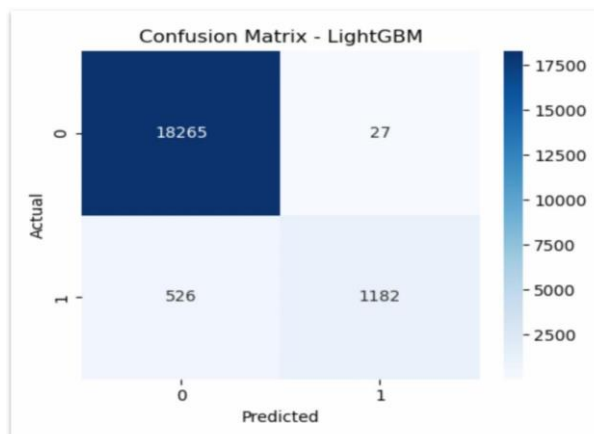


Figure 2(E). Confusion Matrix of Light GBM.

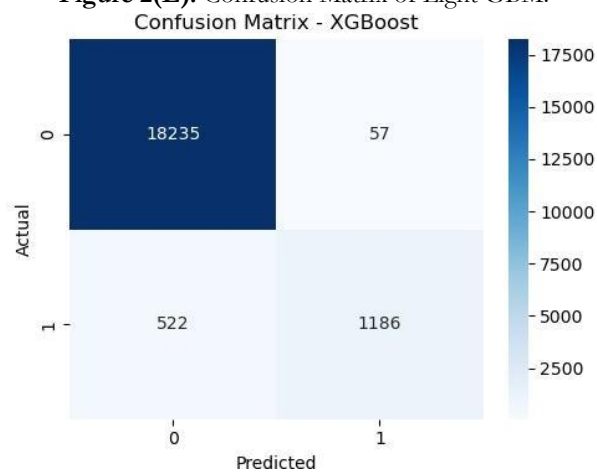


Figure 2(F). Confusion Matrix of XGBoost.

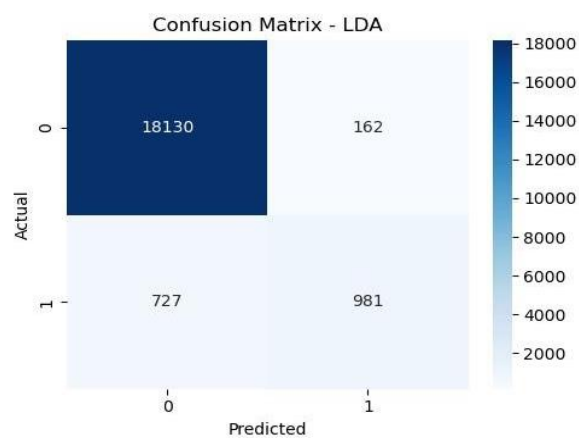


Figure 2(G). Confusion Matrix of LDA.

The Decision Tree, although simple to interpret, performed worse due to its sensitivity to noise and overfitting. SVM performed well but was kernel and regularization parameter tuned sensitive. KNN performed with moderate accuracy but was sensitive to 'k' value and input feature scaling.

These results point towards the need for model selection and preprocessing in designing accurate diagnostic systems. Ensemble methodologies like Random Forest can utilize multiple decision boundaries to increase generalization on novel data.

5 | Comparison with Previous Research Papers

Table 3. Comparison between Present Study and Prior Research on Diabetes Prediction.

Aspect	Paper 1Diabetes Prediction Using Machine Learning	Paper 2Hybrid Machine Learning Models: A Comparative Analysis
Objective	Classify patients into High! / Moderate / Minimal risk based on final score	Predict diabetes presence. (1 or 0) using ML algorithms
Dataset	Medical data from Pune – 2327 cases	Large dataset containing over 100,000 records
Features	Age, weight, RBS, cholesterol, BMI	Gender, age, blood pressure, BMI, cholesterol, HbA1c, smoking, etc.
ML Models	SVM, Decision Tree, Random Forest, KNN	Same models, but hybrid model was used for optimization
Preprocessing	No preprocessing. unbalanced data	Preprocessing: Imputation, Scaling, Encoding, SMOTE, Tuning
Evaluation	Only Accuracy	Accuracy, Precision, Recall, F1-score, Confusion Matrix
Results	Best model: SVM (77.27%)	Best model: Random Forest (97.07%)
Conclusion	SVM was the strongest, especially in early classification	Random Forest outperformed others due to the large dataset and preprocessing; confirmed importance of good model. preparation

Table 4. Previous research papers key findings and used models.

ref	Study Name	Models Used	Key Findings
-	Enhancing the Accuracy of Diabetes Prediction Using Machine Learning Models: A Comparative Analysis	Decision Tree, Random Forest, SVM, KNN	Random Forest best (97.1% accuracy, 0.80 F1-score)
[11]	Diabetes Prediction Using Machine Learning Algorithms	Decision Tree, Random Forest, SVM, KNN	SVM best (77.27% accuracy)
[12]	Prediction of Diabetes Using Neural Network- Based Classifiers	CNN + SAE + VAE	95.14% accuracy

6 | Conclusion

Here, we evaluated and compared the performance of four supervised machine learning algorithms—Decision Tree, Random Forest, Support Vector Machine, and K-Nearest Neighbors—on a real-world large-scale dataset of 100,000 patient records. By incorporating strong preprocessing steps such as imputation, scaling, SMOTE class balancing, and hyperparameter tuning, we achieved a significant improvement in the performance of our model. Of all the models, Random Forest was the top performer with 97.07% accuracy and a strong F1-score of 80%, demonstrating the capability of ensemble methodologies to handle hard, unbalanced health data. These findings underscore the importance of model selection, preprocessing of data, and testing beyond accuracy in robust diagnostic system design. Future work involves the application of deep learning methods such as neural networks and time- series models (e.g., LSTM) and incorporation of clinical EHR data to further enhance prediction capability and real-world applicability in healthcare systems.

Funding

This research has no funding source.

Data Availability

The datasets generated during and/or analyzed during the current study are not publicly available due to the privacy-preserving nature of the data, but are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that there is no conflict of interest in the research.

Ethical Approval

This article does not contain any studies with human participants or animals performed by any of the authors.

References

- [1] World Health Organization. (2016). Global Report on Diabetes. Geneva: World Health Organization. ISBN 978-92-4-156525-7.
- [2] International Diabetes Federation. (2021). IDF Diabetes Atlas (10th ed.). Brussels: International Diabetes Federation. ISBN 978-2-930229-98-0.

- [3] Das, Surajit; Nayak, Samaleswari P.; Sahoo, Biswajit; Nayak, Sarat Chandra. (2024). Machine learning in healthcare analytics: A state-of-the-art review. *Archives of Computational Methods in Engineering*, 31(7), 3923–3962. <https://doi.org/10.1007/s11831-024-10098-3>
- [4] Alshdaifat, Esra'a; Alshdaifat, Doa'a; Alsarhan, Ayoub; Hussein, Fairouz; El-Salhi, Subhieh Moh'd Faraj S. (2021). The effect of preprocessing techniques, applied to numeric features, on classification algorithms' performance. *Data*, 6(2), 11. <https://doi.org/10.3390/data6020011>
- [5] Ganie, Shahid Mohammad; Dutta Pramanik, Pijush Kanti; Malik, Majid Bashir; Mallik, Saurav; Qin, Hong. (2023). An ensemble learning approach for diabetes prediction using boosting techniques. *Frontiers in Genetics*, 14, 1252159. <https://doi.org/10.3389/fgene.2023.1252159>
- [6] Kar, Bikram; Sarkar, Bikash Kanti. (2022). A hybrid feature reduction approach for medical decision support system. *Mathematical Problems in Engineering*, 2022, Article ID 3984082. <https://doi.org/10.1155/2022/3984082>
- [7] Joel, Luke Oluwaseye; Doorsamy, Wesley; Paul, Babu Sena. (2024). On the performance of imputation techniques for missing values on healthcare datasets. *arXiv preprint*, arXiv:2403.14687. <https://arxiv.org/abs/2403.14687>
- [8] Sreejith, S.; Nehemiah, H. Khanna; Kannan, Arputharaj. (2020). Clinical data classification using an enhanced SMOTE and chaotic evolutionary feature selection. *Computers in Biology and Medicine*, 126, 103991. <https://doi.org/10.1016/j.combiomed.2020.103991>
- [9] Awad, Mariette; Khanna, Rahul. (2015). Support Vector Machines for Classification. In *Efficient Learning Machines* (pp. 39–66). Berkeley, CA: Apress. Peterson, Leif E. (2009). K-nearest neighbor. *Scholarpedia*, 4(2), 1883. <https://doi.org/10.4249/scholarpedia.1883>
- [10] Kramer, Oliver. (2016). Scikit-learn. In *Machine Learning for Evolution Strategies* (pp. 45–53). Springer.
- [11] Pewekar, S.; Rathi, S.; Nimje, M. (2024). Diabetes prediction using machine learning algorithms. *Lecture Notes in Networks and Systems* (Vol. 915, pp. 39–46). Springer. https://doi.org/10.1007/978-981-97-7862-1_5
- [12] Kowsher, Md.; Turaba, Mahbuba Yesmin; Sajed, Tanvir; Rahman, M. M. Mahabubur. (2023). Prognosis and treatment prediction of Type-2 diabetes using deep neural network and machine learning classifiers. *arXiv preprint*, arXiv:2301.03093. <https://doi.org/10.48550/arXiv.2301.03093>

Disclaimer/Publisher's Note: The perspectives, opinions, and data shared in all publications are the sole responsibility of the individual authors and contributors, and do not necessarily reflect the views of Sciences Force or the editorial team. Sciences Force and the editorial team disclaim any liability for potential harm to individuals or property resulting from the ideas, methods, instructions, or products referenced in the content.