

Paper Type: Original Article

Applying Supervised Machine Learning to Predict Diabetes from E-Health Records

Ayman H. Abdel-aziem ^{1,*}  and Shimaa Osman ² 

¹ Faculty of Information Systems and Computer Science, October 6th University, Cairo, Egypt;
ayman.hasanein.comp@o6u.edu.eg.

² Computer Department, Higher Institute of Computers and Information Technology, El Shorouk Academy, Cairo, Egypt;
dr.shimaa.osman@sha.edu.eg.

Received: 29 Aug 2025

Revised: 13 Sep 2025

Accepted: 08 Dec 2025

Published: 11 Dec 2025

Abstract

Diabetes mellitus remains a major global health issue, particularly in the United States, where complications from persistent hyperglycemia contribute to severe health outcomes. Although prior research has improved diabetes prediction models, there is still a need for more accurate and efficient machine learning approaches to support early diagnosis and timely intervention. This study explores early diabetes detection using machine learning techniques applied to electronic health records. Four supervised algorithms—Random Forest, Gradient Boosting, Support Vector Classifier (SVC), and K-Nearest Neighbors (KNN) were compared to identifying the most relevant predictive features and determining which model performs best across training, testing, and validation phases. Among the evaluated models, Random Forest and Gradient Boosting demonstrated the strongest and most consistent results, achieving training accuracy of approximately 82.5% and test accuracy around 82.2%. Although the KNN model showed high sensitivity (98.54%) and specificity (93.63%) in a separate assessment, its test accuracy in this study was slightly lower at 77.9%. The SVC model reached a test accuracy of 76.7%. Validation outcomes further confirmed the stability and generalizability of the Random Forest and Gradient Boosting models when applied to unseen data. Overall, the results indicate that ensemble methods deliver more accurate and reliable predictions for diabetes diagnosis compared to KNN and SVC, making them more suitable for clinical settings. While KNN performed well under certain conditions, ensemble techniques provided better generalization. Future work should expand the feature set and apply thorough hyperparameter optimization to improve predictive performance and enable more effective early detection and management of diabetes.

Keywords: Diabetes Mellitus, Machine Learning, Early Diagnosis, Predictive Models, Random Forest, Gradient Boosting, K-Nearest Neighbors, Clinical Use.

1 | Introduction

Diabetes is a chronic metabolic disease characterized by elevated levels of blood glucose that, over time, can seriously damage the heart, blood vessels, eyes, kidneys, and nerves [1]. The most general form is Type 2 diabetes, found in adults [2], and the condition develops when muscles become resistant to insulin or when the body does not produce enough of it—a condition whose global prevalence has risen dramatically in recent decades. In contrast, Type 1 diabetes is a chronic condition where the pancreas produces little or no insulin



Corresponding Author: ayman.hasanein.comp@o6u.edu.eg



<https://doi.org/10.61356/j.scin.2025.2624>



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>).

[3]. For all individuals living with this condition, reliable and affordable access to treatment, especially insulin, is essential for survival, while a global target has also been set to halt the rise in both diabetes and obesity until 2025 [4]. An estimated 830 million people worldwide are affected by diabetes, and most of them come from low- and middle-income countries [5]. More than half of those affected do not receive proper treatment. Both the number of people diagnosed with diabetes and the number of untreated cases has increased steadily over recent decades. The proposed approach focuses on predicting diabetes using techniques of machine learning. Steps to implement the proposed system are given, then its performance is evaluated by applying two widely used machine learning algorithms on two different real-life diabetes datasets. Before presenting the system, in Section 2 an overview of related research is presented concerning diabetes prediction based on machine learning. In Section 3 the SMOTE (Synthetic Minority Over-Sampling Technique) is introduced as an oversampling method to handle class imbalance by generating artificial samples of the minority class. Numerical experiments and comparative performance results are given in Section 4. Finally, Section 5 concludes the paper and includes also directions for further research.

2 | Related Work

Studies on diabetes prediction using advanced machine learning techniques are, therefore, well documented, showing the different methodologies that have been used in these various studies. This comparative study shows significant heterogeneity in the data sources and algorithmic approaches used, besides being committed to rigorous performance evaluation. Thus, several studies used highly targeted private datasets, such as electronic health records obtained from hospitals in Saudi Arabia [6], specialized primary healthcare institutions in Slovenia [7], and a private medical foundation located in Seoul, South Korea [8]. Such geographically and institutionally specific data usage indicates a belief in the best model generalization and clinical relevance locally. Other researchers, on the other hand, utilized well-established, sometimes public, large-scale survey datasets, such as the widely recognized PIMA Indian dataset [9, 10, 11], which gives a uniform comparison benchmark. The other large-scale datasets included the National Health and Nutrition Examination Survey (NHANES) [9], data from an important Henan rural cohort study in China [12], and patient records provided by a CBHS health funds company in Australia [13]. The diversity in datasets indicates the challenge and necessity of developing models with good generalization ability over different populations and clinical settings.

This is equally varied at the algorithmic level, containing both classic statistical and powerful non-linear models. Classic approaches generally rest on standard techniques such as Logistic Regression (LR), Support Vector Machines (SVM), and tree-based approaches like Decision Trees (DT) and Random Forests (RF). Yet, many of the recent studies have used sophisticated algorithms to model complex interactions in the data. For instance, high-performance, tree-based gradient boosting frameworks such as XGBoost [7, 8, 13] and Light GBM [7] have been used, which have been known for their predictive accuracy and efficiency. Another important direction of research pertains to the use of ensemble techniques. Explicit Majority Voting (EMV) [6], soft voting, stacking, and confusion matrix-based ensembles [8] have been studied in the literature for combining the predictions of multiple base classifiers with the aim of improving model robustness and stability. The use of Artificial Neural Networks ANN, k-Nearest Neighbors (kNN), and methods such as Naive Bayes NB [11] further demonstrates the wide range of machine learning techniques tested so far to achieve the best possible accuracy.

Finally, model validation depends on a wide range of evaluation metrics to be able to offer an overview of its performance from different angles. The main standard set of metrics applied, which together characterize the balance between correctness and completeness of positive predictions, includes Accuracy, precision, recall (or sensitivity), and the F-measure [6, 8, 9, 11, 13]. Also, with regard to binary classification tasks, important metrics are the AUC [7, 9, 11, 12, 13]-i.e., the Area Under the Receiver Operating Characteristic Curve-which depicts the model's ability to correctly distinguish diabetic and non-diabetic cases at all classification thresholds. Other specialized metrics were also reported, such as the Root Mean Square Error-RMSE-[7], Matthews Correlation Coefficient-MCC-[8], along with statistical fit indices like Akaike Information

Criterion-AIC-and Bayesian Information Criterion-BIC-[10], ensuring a thorough statistical and clinical appraisal of the machine learning systems proposed. This rigorous evaluation method ensures that the progress made by ML-based diabetes prediction is statistically sound and clinically meaningful.

Table 1. Diabetes dataset of variables.

	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree unction	Age	Outcome
Pregnancies	1	0.129459	0.141282	- 0.081672	- 0.073535	0.017683	- 0.033523	0.544341	0.221898
Glucose	0.129459	1	0.152590	0.057328	0.331357	0.221071	0.137337	0.263514	0.466581
Blood Pressure	0.141282	0.152590	1	0.207371	0.088933	0.281805	0.041265	0.239528	0.065068
Skin Thickness	-0.081672	0.057328	0.207371	1	0.436783	0.392573	0.183928	- 0.113970	0.074752
Insulin	-0.073535	0.331357	0.088933	0.436783	1	0.197859	0.185071	- 0.042163	0.130548
BMI	0.017683	0.221071	0.281805	0.392573	0.197859	1	0.140647	0.036242	0.292695
Diabetes Pedigree Function	-0.033523	0.137337	0.041265	0.183928	0.185071	0.140647	1	0.033561	0.173844
Age	0.544341	0.263514	0.239528	- 0.113970	- 0.042163	0.036242	0.033561	1	0.238356
Outcome	0.221898	0.466581	0.065068	0.074752	0.130548	0.292695	0.173844	0.238356	1

SMOTE: Synthetic Minority Oversampling Technique: Imbalanced classification refers to the development of predictive models on datasets in which one class (often the class of interest) is significantly underrepresented. This imbalance presents a challenge because most machine learning algorithms tend to bias their predictions toward the majority class, resulting in deficient performance in the minority class. However, in many applications such as medical diagnosis the accurate prediction of the minority class is of critical importance. One common strategy for addressing this issue is oversampling the minority class [14–16]. A simple form of oversampling involves duplicating existing minority-class instances, but this approach does not introduce added information and may increase the risk of overfitting. To overcome this limitation, synthetic examples can be generated from existing samples. This process, known as the Synthetic Minority Oversampling Technique (SMOTE), creates new minority-class instances by interpolating between existing ones. SMOTE is therefore considered a data augmentation technique that enhances the representation of the minority class without merely replicating data. Figure 1 illustrates the class distribution before and after applying SMOTE, demonstrating its effect on balancing the dataset. Figure 2 provides an algorithmic flowchart summarizing the key steps involved in the SMOTE procedure.

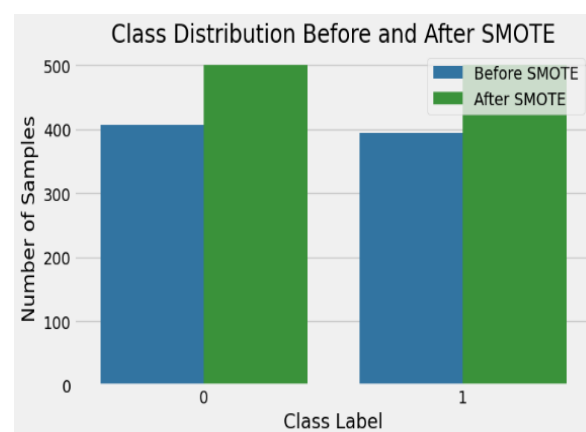
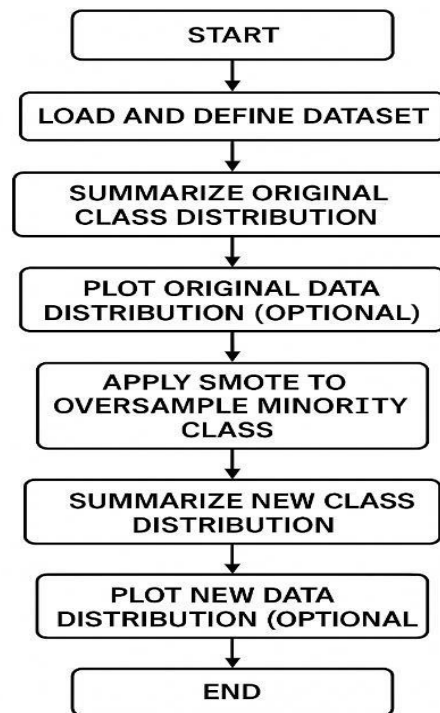


Figure 1. Class distribution before & after SMOTE.



Algorithm 1: SMOTE Preprocessing Steps (Flowchart)

Figure 2. Steps of Applying SMOTE to Balance an Imbalanced Dataset.

SMOTE algorithm starts with reading and loading the dataset, identifying minority class samples, identifying k-nearest neighbors for every sample, selecting one or more neighbors randomly, creating synthetic samples between the sample and neighbors, adding new samples to the dataset, and finally training using the new balanced dataset.

3 | Experimental Environment

Comparison of Machine Learning Algorithms

This section presents a comparative analysis of four widely used supervised machine learning algorithms for diabetes prediction. The selected models—K-Nearest Neighbors (KNN), Random Forest (RF), Gradient Boosting (GB), and Support Vector Classifier (SVC)—have been frequently applied in previous studies due to their effectiveness in handling medical classification tasks.

The evaluation is conducted using a publicly available diabetes dataset obtained from Kaggle, a widely used online repository known for its extensive collection of high-quality datasets. The dataset, provided in CSV format, is used to train and test each of the selected algorithms under identical conditions. Model performance is assessed using standard evaluation metrics, including accuracy, precision, recall, and F1-score, which collectively provide a comprehensive measure of predictive capability.

The algorithms considered in this study are (Random Forest (RF) - Gradient Boosting Classifier (GB)- Support Vector Classifier (SVC) - K-Nearest Neighbors (KNN))

Each model is applied to the dataset to determine its effectiveness in predicting diabetes outcomes. Through this comparative analysis, we aim to identify which algorithm demonstrates superior performance and is therefore most suitable for diabetes prediction based on the given dataset.

Table 2. Feature and Class Distribution of the Diabetes Dataset.

Dataset	No. of features	Features	Number of observations in diabetic class	Number of observations in non-diabetic class	Total number of observation
Diabetic patients	8	The features are Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, age Outcome.	268 (34.9%)	500 (65.1%)	768

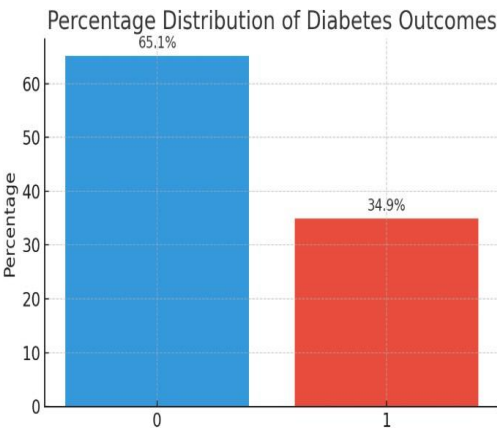


Figure 3. Percentage Distribution of Diabetes Outcome.

3.1 | Demonstrates

Figure 4 illustrates which characteristics—such as age, blood pressure, glucose level, and others—are correlated with the presence or absence of diabetes in the given dataset. A higher positive or negative correlation value indicates a stronger association between the characteristic and the diabetes outcome. For instance, if glucose level exhibits a high positive correlation, this suggests that individuals with elevated glucose levels are more likely to be diabetic. This type of analysis is valuable because it highlights the most significant predictors of diabetes, which can subsequently be used to develop more accurate and effective prediction models.

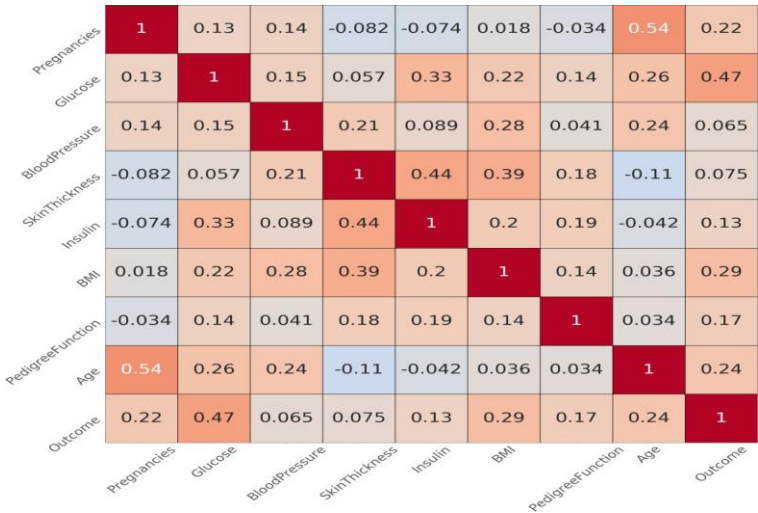


Figure 4. Correlation between features and diabetic/non-diabetic class for dataset that be used.

3.2 | Box Plots

Box plots also known as box-and-whisker plots are straightforward yet powerful data visualizations. Their simplicity is, in fact, one of their key advantages, as they can communicate several summary statistics quickly and efficiently. Typically, when examining the distribution of a variable (for example, miles per gallon (MPG) across different car types), one would need to plot each MPG value against each car category individually. Such visualizations occupy considerable space and make comparisons across multiple independent variables more difficult. For instance, if histograms were used to compare MPG across car styles, a separate histogram would need to be created for each category, followed by a manual comparison among all car styles—an approach that is both time-consuming and visually complex.

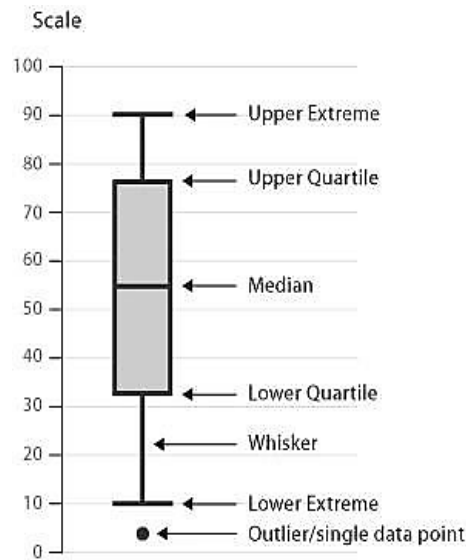


Figure 5. Percentage Distribution of Diabetes.

Explaining Statistical Measures through a Box Plot

- **Upper Extreme:**
Represents the highest data point that falls within the normal range (i.e., not classified as an outlier). It indicates the maximum non-outlier value in the dataset.
- **Upper Quartile (Q3):**
Corresponds to the 75th percentile. This means that 75% of the data values are below this point.
- **Median (Q2):**
The middle value of the dataset, representing the 50th percentile. It divides the data distribution into two equal halves.
- **Lower Quartile (Q1):**
Represents the 25th percentile, indicating that 25% of the data values lie below this point.
- **Whiskers:**
The lines extending from the box toward the lowest and highest data values that are not considered outliers.
- **Lower Extreme:**

- The lowest data point within the acceptable range (not an outlier). It marks the minimum non-outlier value in the dataset.
- Outlier / Single Data Point:
- A value that deviates significantly from the overall distribution pattern. It is typically represented as an isolated dot (●).

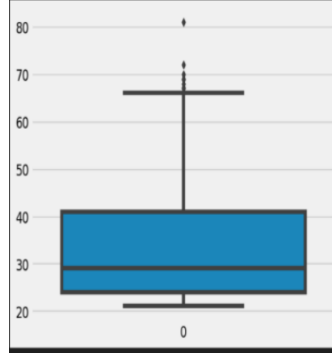


Figure 6. Box Plot.

4 | Results and Discussions

We now proceed to present the results obtained from the four machine learning models. Among these models, Random Forest and Gradient Boosting demonstrated the highest levels of accuracy and consistency, achieving the strongest performance scores. The performance of all models will be evaluated using several key metrics and supported by a range of visualizations (graphs and plots).

In this analysis, particular attention is given to the most significant performance measures: Accuracy, Recall, and F1-Score. Accuracy reflects the overall correctness of the model's predictions; Recall indicates the model's ability to correctly identify all diabetic cases; and the F1-Score provides a balanced measure that combines both precision and recall—making it especially valuable when working with imbalanced datasets.

4.1 | Variables and Equations

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

$$\text{Precision} = \frac{TP + TN}{TP + FP} \quad (2)$$

$$\text{Recall} = \frac{TP + TN}{TP + FN} \quad (3)$$

$$\text{F - Measure} = \frac{2(\text{recall} * \text{precision})}{\text{recall} + \text{precision}} \quad (4)$$

The dataset was split into 70% for training and 30% for testing, ensuring that the models could be evaluated rigorously and their performance assessed with utmost care.

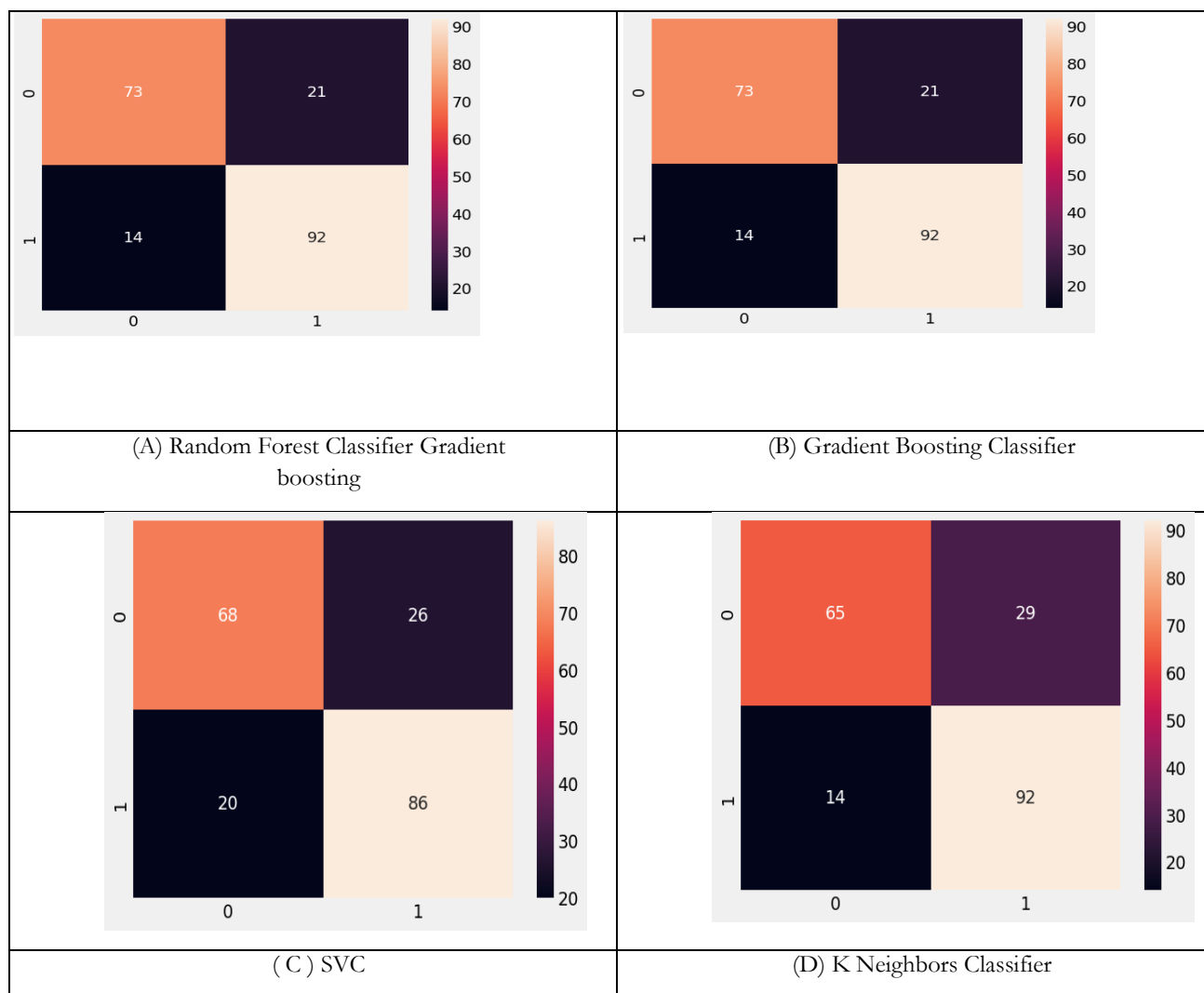


Figure 7. Different Algorithm Types Random Forest

And finally, after all the details and figures, we understand what the algorithms look like and where each one stands.

Table 3. Performance Metrics of Machine Learning Algorithms.

Algorithms	Accuracies	Recall score	F1 score
Random Forest Classifier	0.825	0.822260	0.823406
Gradient Boosting Classifier	0.825	0.822260	0.823406
SVC	0.770	0.767363	0.768122
K Neighbors Classifier	0.785	0.779707	0.781009

5 | Conclusion

This research has covered the effectiveness of several machine learning models supervised-mainly Random Forest, Gradient Boosting, Support Vector Classifier, and K-Nearest Neighbors-for early prediction of diabetes utilizing electronic health records. The main results clearly indicated the superiority and stability of the performance of ensemble methods. Both Random Forest and Gradient Boosting showed the best results, with the same training accuracy of about 82.5% and a test accuracy of about 82.2% with great stability. These high and stable performances confirm the dependability of both methods and their generalization on unseen

data. The study also confirmed the importance of Glucose, BMI, and Age as significant features and the most powerful predictors of the outcome of diabetes, and the important use of SMOTE in addressing the class imbalance problem of the dataset contributed much to the reliability of the models. With regard to model comparison, the best overall model had a test accuracy of 77.9%, the second was KNN, while the model that proved the weakest model is SVC at an accuracy level of 76.7%, meaning although individual non-ensemble models perform well for a particular metric, the performance of these two modeling ensembles techniques gives better overall generalization and is suitable for clinical risk assessment. In view of this, ensemble methods, such as Random Forest and Gradient Boosting, provide the most accurate and reliable results for the diagnosis of diabetes, which will be suitably applied to a clinical decision-support system to allow more effective early detection and management of the patient.

Future work should focus on:

- **Feature Expansion:** Incorporating a broader range of real-world clinical characteristics and biomarkers into the feature set.
- **Advanced Techniques:** Exploring more advanced deep learning techniques and thorough hyperparameter optimization to further enhance the predictive accuracy and diagnostic performance of the models.

Funding

This research has no funding source.

Data Availability

The datasets used and analyzed in this study are publicly available. The primary data sources include the PIMA Indian and Sylhet datasets, supplemented with additional data from open-access resources on Kaggle. These datasets are anonymized and do not involve any privacy or ethical concerns, making them appropriate for open academic research.

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, "Diabetes is a chronic disease," WHO, Accessed: Jan. 2025. [Online]. Available: https://www.who.int/health-topics/diabetes#tab=tab_1
- [2] World Health Organization, "The most common is type 2 diabetes, usually in adults," WHO, Accessed: Jan. 2025. [Online]. Available: https://www.who.int/health-topics/diabetes#tab=tab_1
- [3] World Health Organization, "Access to affordable treatment for people living with diabetes," WHO, Accessed: Jan. 2025. [Online]. Available: https://www.who.int/health-topics/diabetes#tab=tab_1
- [4] World Health Organization, "About 830 million people worldwide have diabetes," WHO, Accessed: Jan. 2025. [Online]. Available: https://www.who.int/health-topics/diabetes#tab=tab_1
- [5] World Health Organization, "More than half of people living with diabetes are not receiving treatment," WHO, Accessed: Jan. 2025. [Online]. Available: https://www.who.int/health-topics/diabetes#tab=tab_1
- [6] S. Montaser, "Investigating health-related features and their impact on the prediction of diabetes using machine learning," *Applied Sciences*, 2021. doi: 10.3390/app11031173.
- [7] S. Montaser, "Early detection of type 2 diabetes mellitus using machine learning-based prediction models," *Scientific Reports*, 2020. doi: 10.1038/s41598-020-68771-z.

- [8] S. Montaser, "Prediction of type 2 diabetes based on machine learning algorithm," *Int. J. Environ. Res. Public Health*, 2021. doi: 10.3390/ijerph18063317.
- [9] S. Montaser, "Machine learning–based application for predicting risk of type 2 diabetes mellitus (T2DM) in Saudi Arabia: A retrospective cross-sectional study," *IEEE Access*, 2020. doi: 10.1109/ACCESS.2020.3035026.
- [10] S. Montaser, "Predicting type 2 diabetes using logistic regression and machine learning approaches," *Int. J. Environ. Res. Public Health*, 2021. doi: 10.3390/ijerph18147346.
- [11] S. Montaser, "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms," *Neural Computing and Applications*, 2022. doi: 10.1007/s00521-022-07049-z.
- [12] S. Montaser, "Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: The Henan Rural Cohort Study," *Scientific Reports*, 2020. doi: 10.1038/s41598-020-61123-x.
- [13] S. Montaser, "A patient network-based machine learning model for disease prediction: The case of type 2 diabetes mellitus," *Applied Intelligence*, 2022. doi: 10.1007/s10489-021-02533-w.
- [14] I. H. Witten, E. Frank, and M. A. Hall, "Challenges of working with imbalanced datasets," in *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed., Springer, 2016. doi: 10.1007/978-981-10-1741-4.
- [15] J. Brownlee, "The simplest approach involves duplicating minority-class examples," *Machine Learning Mastery*, Accessed: Jan. 2025. [Online]. Available: <https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/>
- [16] J. Brownlee, "Synthetic Minority Oversampling Technique (SMOTE)," *Machine Learning Mastery*, Accessed: Jan. 2025. [Online]. Available: <https://machinelearningmastery.com/how-to-configure-image-data-augmentation-in-keras/>

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.