Journal of Science and Technology – Smart Systems and Devices
Volume 36, Issue 3, Month 2026, 000–000

Early Detection of Diabetes Using Machine Learning

Tran Anh Vu[1], Pham Thi Hong Tham[2], Vu Dan Vy[1], Hoang Quang Huy[1], Pham Thi Viet Huong[2,*]
1 Hanoi University of Science and Technology, Ha Noi, Vietnam
2 International School, Vietnam National University, Ha Noi, Vietnam
** Corresponding author email: huongpv@vnu.edu.vn*

Abstract

Early detection of diabetes mellitus is essential for mitigating severe complications such as renal failure and cardiovascular disease. This study presents an optimized machine learning framework for the early diagnosis of diabetes using the Pima Indians Diabetes dataset. To address inherent data quality issues, we implemented a rigorous preprocessing pipeline comprising median-based missing value imputation, interquartile range (IQR) outlier removal, and the Synthetic Minority Over-sampling Technique (SMOTE) to rectify class imbalance. We evaluated two ensemble architectures, Random Forest (RF) and XGBoost, integrated with Grid Search for systematic hyperparameter optimization. Experimental results across three scenarios demonstrated that feature selection significantly impacts predictive integrity; specifically, maintaining Insulin as a core feature while strategically excluding Skin Thickness enhanced model stability. The Random Forest model achieved a peak accuracy of 96.61% with a near-perfect recall rate, while XGBoost reached 95.76% accuracy. By outperforming several contemporary models, this research underscores the necessity of synergistic data preprocessing and ensemble learning in clinical diagnostics. These findings provide a robust decision-support tool for healthcare providers to facilitate timely intervention and improved patient outcomes.

Keywords: Diabetes, early detection, grid search, machine learning, SMOTE.

1. Introduction

Diabetes mellitus results from insufficient insulin production by the pancreas or impaired insulin function due to acquired or genetic factors, leading to elevated blood glucose levels [1]. Disease-related complications include blindness, renal failure, cardiovascular diseases, and even death [2]. However, effective glycemic control through a healthy diet, regular physical activity, and adherence to physician-prescribed medication can help prevent or delay these complications. Currently, unhealthy lifestyles have contributed to an increasing prevalence of conditions such as diabetes, obesity, and hypertension. Therefore, early detection of diabetes plays a crucial role in disease prevention, enabling timely intervention to prevent complications and reduce the risk of transmission to future generations.

In recent years, the academic community has witnessed a surge in research dedicated to the early diagnosis of diabetes through machine learning [3, 4, 5] and deep learning architectures [6, 7]. These computational advancements have fundamentally improved diagnostic precision, facilitating timely interventions and more robust patient management strategies. For instance, Rastogi [8] evaluated several machine learning techniques; in this study, Logistic Regression emerged as the most effective model with an accuracy of 82.46%. Concurrently, Reza et al. [4] enhanced SVM performance by implementing a nonlinear kernel, which achieved an 85.5% accuracy rate for type 2 diabetes classification, thereby surpassing traditional kernel methods and demonstrating significant potential for clinical screening.

Furthermore, research by another group [9] integrated Artificial Neural Networks (ANNs) with sophisticated data processing techniques to minimize predictive error margins. Saeed [10] later identified Random Forest as a superior algorithm, reaching an accuracy of 94.2%, whereas SVM and Logistic Regression showed only moderate efficacy. This high performance of Random Forest was corroborated by Thotad [3], who reported a near-perfect classification accuracy of 99.87%. Beyond these methods, the integration of XGBoost in recent literature [5, 11] has markedly improved the identification of at-risk individuals who lack overt clinical symptoms.

The evolution of deep learning is further evidenced in study [12], where Convolutional Neural Networks (CNNs) were coupled with Bayesian optimization. By employing the SMOTE technique for class balancing and automated hyperparameter tuning, this model achieved 89.36% accuracy, outperforming several conventional algorithms. Similarly, the utility of K-Nearest Neighbors (KNN) was validated by research [13] using the Pima Indians Diabetes Dataset, proving its effectiveness in forecasting disease onset from clinical variables. Expanding on ensemble techniques,

p-ISSN 3093-3285
e-ISSN 3093-3315
https://doi.org/10.51316/jst.192.ssad.2026.36.3.xx
Received: Jan 16, 2026; Revised: Apr 3, 2026;
Accepted: Mar 29, 2026; Online: Jun 24. 2026

Hasan [14] analyzed multiple algorithms–including AdaBoost, XGBoost, and Multi-Layer Perceptron–finding that ensemble-based models achieved superior results with an AUC of 0.950 and a specificity of 0.934.

Recent architectural innovations include a model based on Convolutional Long Short-Term Memory (CLSTM) [6], which utilized multivariate time-series interpolation for enhanced preprocessing; the results demonstrated its superiority over traditional machine learning approaches on the PIDD dataset. Finally, in a comparative study conducted in Iran, Heydari et al. [15] determined that ANNs reached a peak accuracy of 97.44%. By mimicking biological neural systems through multi-layered interconnections, ANNs showcased data processing capabilities that exceed those of rival algorithms. Collectively, these milestones underscore the transformative potential of machine and deep learning in modern diabetes management and clinical diagnostics.

This study proposes the implementation of Random Forest and XGBoost algorithms to facilitate diabetes prediction and the development of proactive diagnostic strategies. By leveraging these ensemble methods, the proposed system generates precise clinical insights, optimizes predictive accuracy, and enhances the reliability of early-stage diabetes mellitus detection, thereby supporting more effective patient health management.

2. Materials and Methods

This study is constructed upon a rigorous data analysis workflow, extending from initial data acquisition to the achievement of optimal predictive models. The main purpose is to ensure the precision and generalizability in the early detection of diabetes mellitus.

The methodological structure is divided into three distinct phases:

- Dataset Description: Utilization of the Pima Indians Diabetes dataset to delineate key biological attributes that influence the disease's pathology.

- Data Preprocessing: Implementation of a comprehensive pipeline encompassing missing value imputation, outlier detection, and the application of Synthetic Minority Over–sampling Technique (SMOTE) to mitigate class imbalance.

- Model Developmentand Optimization: Employment of robustensemble algorithms–specifically Random Forest and XGBoost–integrated with Grid Search for hyperparameter tuning to determine the most effective model configurations.

The experimental process employs a Stratified K-Fold cross-validation scheme to maintain result stability and objectivity across heterogeneous data subsets. By utilizing this framework, the study minimizes systematic error and validates the robustness of the model's performance.

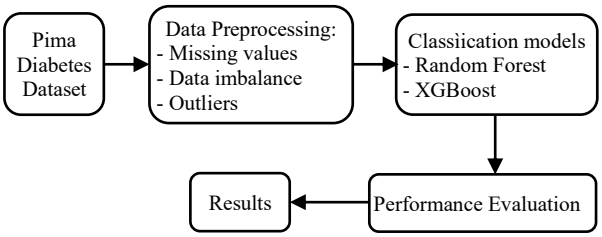

Fig. 1. The workflow of the system

2.1 Dataset Description

This research utilizes the Pima Indians Diabetes dataset, originally collected in Phoenix, Arizona, involving subjects aged 21 years and older. This publicly accessible dataset is a benchmark in the field, having been extensively validated in prior literature [4, 13, 16]. The cohort consists of 768 female patients characterized by eight clinical features: Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function (DPF), and Age. The target variable is the diagnostic "Outcome." Comprehensive variable definitions are provided in Table 1, while Table 2 presents the initial ten records of the dataset. The class distribution of the target variable is visualized in the pie chart in Fig. 2. This dataset was sourced via Kaggle under the CC0 Public Domain License.

The Pima Indians Diabetes dataset consists of 768 total samples across eight independent biological attributes and one target variable. All variables in the dataset are represented as integer data types.

The Pima Indians Diabetes dataset comprises eight physiological attributes used as predictors for the diagnostic outcome. The specific characteristics and ranges of these variables are detailed below:

- Pregnancies: Quantifies the gravidity of the patient, representing the total number of pregnancies, with observed values ranging from 0 to 17.

- Glucose: Measures the plasma glucose concentration (mg/dL) two hours after an oral glucose tolerance test (OGTT). The recorded values range from 0 to 199.

- Blood Pressure (BP): Refers specifically to the diastolic blood pressure measured in millimeters of mercury (mmHg), with a recorded range of 0 to 122.

- Skin Thickness (ST): Represents the triceps skinfold thickness measured in millimeters (mm), serving as an indicator of subcutaneous fat, with values between 0 and 99.

- Insulin: Indicates the serum insulin levels (μU/mL) two hours post-glucose administration, ranging from 0 to 846.

- Body Mass Index (BMI): A derived metric of body fatness calculated as weight in kilograms divided by the square of height in meters.

- Diabetes Pedigree Function (DPF): A specialized score that quantifies the hereditary risk of diabetes by incorporating the family history of the disease. Values in this cohort range from 0.078 to 2.42.

- Age: The chronological age of the female subjects, ranging from 21 to 81 years.

- Outcome (Target Variable): A binary classification where 0 denotes a negative diagnosis (non-diabetic) and 1 denotes a positive diagnosis (diabetic).

However, a critical observation of the data shows the presence of missing or null values in the raw data:

- High Completion Variables: Both Age, DPF (Diabetes Pedigree Function), and the Outcome variable are fully populated with 768 records.

- Minor Missing Data: Attributes such as Glucose (763), BMI (757), and Blood Pressure (733) show relatively high completion rates.

- Significant Missing Data: In contrast, Skin Thickness (541) and Insulin (427) exhibit substantial missingness, with the latter containing valid data for only approximately 55.6% of the total samples.

This variance in data density across attributes necessitates the rigorous preprocessing pipeline mentioned previously, specifically the implementation of missing value imputation to ensure model stability and predictive integrity

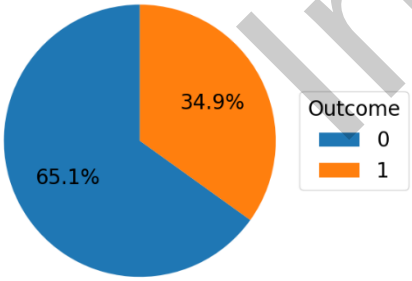

Fig. 2. Pie chart of the data output (0: Number of individuals diagnosed as non-diabetic; 1: Number of individuals diagnosed with diabetes)

2.2. Data Preprocessing

The proposed data preprocessing workflow comprises three main procedures: missing value imputation, outlier removal, and dataset balancing. As shown in Fig. 3, the BMI, Skin Thickness, Insulin, Blood Pressure, and Glucose attributes contain missing values (encoded as zeros), among which Insulin exhibits the highest missing rate of 48.7%, followed by Skin Thickness at 29.56%. As both variables are considered important for diabetes prediction, three missing-value handling scenarios are designed as follows:

Scenario 1: all variables are retained, and missing values are replaced with the median of the corresponding variable.

Scenario 2: the Skin Thickness variable is removed, and the remaining missing values (BMI, Insulin, Blood Pressure, and Glucose) are imputed using the median

Scenario 3: the Insulin variable is removed, and the remaining missing values (BMI, Skin Thickness, Blood Pressure, and Glucose) are imputed using the median.

After missing-value processing under each scenario, outlier removal is performed to enhance predictive accuracy. Outliers are identified using the interquartile range (IQR) method, revealing that all attributes, except Glucose, contain outliers (Fig. 4).

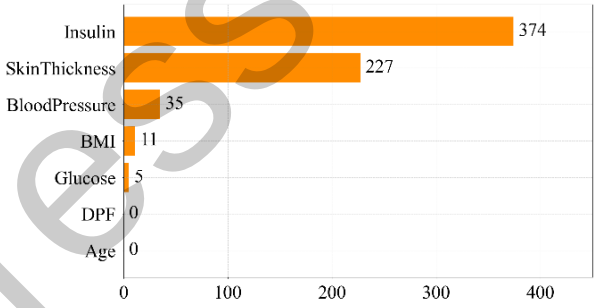

Fig. 3. Number of missing values in the PIMA diabetes dataset

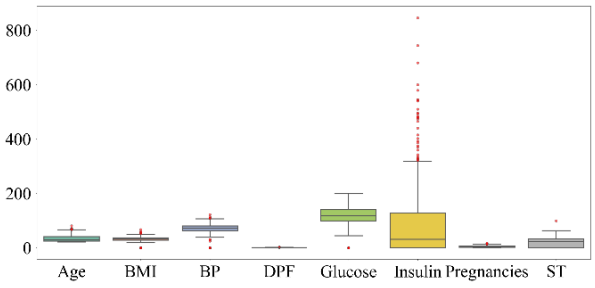

Fig. 4. Outlier values in the PIMA diabetes dataset

The dataset exhibits an inherent class imbalance, which can potentially introduce bias by causing the model to favor the majority class during the learning process. To mitigate this concern, the Synthetic Minority Over-sampling Technique (SMOTE) is implemented to generate synthetic instances for the minority class, thereby achieving a balanced class distribution [17]. The training and testing set are divided using 80:20 ratio to rigorously assess the model's generalization capability on previously unseen data. To further ensure a reliable and robust performance evaluation, a 10-fold Stratified K-Fold cross-validation approach is adopted. Within this

framework, each iteration utilizes one fold for testing while the remaining nine folds are dedicated to training; the stratification process ensures that the original class proportions are strictly preserved across all partitions [18].

2.3. Classification Algorithms

2.3.1. Random forest

The Random Forest (RF) classifier operates as an ensemble learning architecture composed of multiple decision trees, where each constituent tree is independently trained on a distinct bootstrap sample of the dataset. To enhance predictive precision and minimize variance, the RF algorithm aggregates the individual outputs of these trees through mechanisms such as averaging (for regression) or majority voting (for classification). Rather than relying on the outcome of a solitary decision tree, the algorithm synthesizes the collective results from the entire forest to determine the final class label during the prediction phase. Structurally, each internal node within a tree functions as a decision threshold based on a specific feature, strategically partitioning the data into increasingly homogeneous subsets for subsequent analysis.

2.3.2. XGBoost

XGBoost incrementally builds multiple weak learners and combines them into a strong ensemble model [20]. Specifically, after an initial model is trained, the algorithm computes the residuals (i.e., the errors between the actual and predicted values) at each training iteration. A new model is then trained to accurately predict these residuals, and its output is added to the cumulative predictions of the previous models. Through this iterative optimization process, XGBoost achieves a favorable balance between bias and variance, resulting in more stable and robust performance compared with conventional boosting methods.

2.3.3. Grid search

Grid Search is an extensively utilized technique within the machine learning domain for systematic hyperparameter optimization. The optimal model is identified by exhaustively evaluating all possible combinations of predefined hyperparameter values. Specifically, Grid Search constructs a “grid” of candidate values for each hyperparameter (e.g x_1,\ x_2,\ x_3,\ldots.) and systematically trains and evaluates the model across all combinations. At each step, only one hyperparameter configuration is varied while the others are held constant, allowing the isolated assessment of each hyperparameter’s impact [21]. Although Grid Search can effectively fine-tune models to achieve improved predictive performance, its computational efficiency is strongly influenced by dataset size and the number of hyperparameters, making it less suitable for large-scale and high-dimensional datasets.

3. Experimental Procedure

3.1. Data Visualization

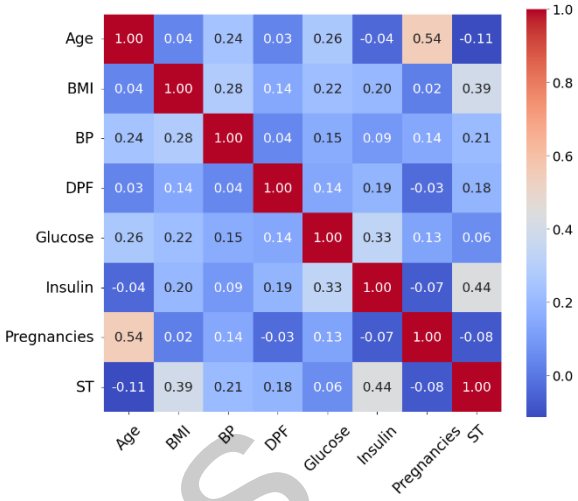

Fig. 5. Correlation matrix before preprocessing

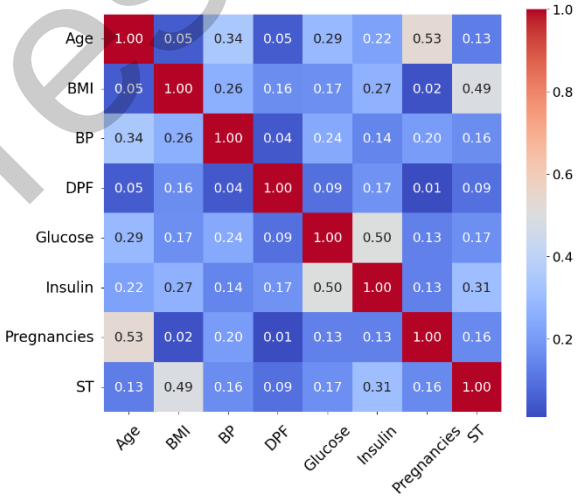

Fig. 6. Correlation matrix after preprocessing

At the initial stage, data visualization is performed to clearly examine the distribution and relationships among variables. Correlation heatmaps before preprocessing (Fig. 5) and after preprocessing (Fig. 6) are generated to analyze and compare the relationships between variables.Prior to preprocessing, both Insulin and Skin Thickness exhibit weak correlations with the Outcome variable. After data preprocessing, the correlation coefficient between Insulin and Outcome increases significantly, indicating that Insulin contains substantial informative value and that removing this variable may reduce model performance. In contrast, Skin Thickness continues to show a weak correlation with Outcome after preprocessing, suggesting that this variable contributes less to the prediction task and may be considered for removal when dimensionality reduction is required.

3.2. Results

Scenario 1 (Table 1): All variables are retained, missing values are replaced with the median, outliers are handled, and the dataset is balanced.

Scenario 2 (Table 2): The Skin Thickness variable is removed, the remaining missing values (BMI, Insulin, Blood Pressure, and Glucose) are imputed using the median, outliers are handled, and the dataset is balanced.

Table 1. Results of Scenario 1

	Model	Acc	Pre	Recall	F1
K-Fold	RF	**0.9136**	**0.8849**	**0.8437**	**0.8618**
	SMOTE_RF	0.8966	0.8239	0.8745	0.8451
	XGB	0.9085	0.8771	0.8384	0.8548
	SMOTE_XGB	0.9034	0.8381	0.8747	0.8535
80/20	RF	0.9407	0.8974	0.9211	0.9091
	SMOTE_RF	**0.9661**	**0.9048**	**1.000**	**0.9500**
	XGB	0.9492	0.9211	0.9211	0.9211
	SMOTE_XGB	0.9492	0.9211	0.9211	0.9211

Table 2. Results of Scenario 2

	Model	Acc	Pre	Recall	F1
K-Fold	RF	0.9169	0.8983	0.8434	0.8675
	SMOTE_RF	0.9034	0.8384	0.8800	0.8551
	XGB	**0.9186**	**0.8951**	**0.8542**	**0.8718**
	SMOTE_XGB	0.8983	0.8285	0.8750	0.8479
80/20	RF	0.9492	0.9444	0.8947	0.9189
	SMOTE_RF	0.9322	0.8571	0.9474	0.9000
	XGB	**0.9576**	**0.9459**	**0.9211**	**0.9333**
	SMOTE_XGB	0.9407	0.8780	0.9474	0.9114

Table 3. Results of Scenario 3

	Model	Acc	Pre	Recall	F1
K-Fold	RF	0.8508	0.8091	0.7124	0.7546
	SMOTE_RF	0.8271	0.7215	0.7800	0.7449
	XGB	0.8678	0.8194	0.7584	0.7827
	SMOTE_XGB	0.8356	0.7175	0.8324	0.7665
80/20	RF	0.8898	0.8378	0.8158	0.8267
	SMOTE_RF	0.9068	0.8140	0.9211	0.8642
	XGB	0.9153	0.8684	0.8684	0.8684
	SMOTE_XGB	0.9322	0.8571	0.9474	0.9000

Scenario 3 (Table 3): The Insulin variable is removed, the remaining missing values (BMI, Skin Thickness, Blood Pressure, and Glucose) are imputed using the median, outliers are handled, and the dataset is balanced.

Overall, the Stratified K-Fold approach yields more stable results, as it trains and evaluates the model across multiple subsets of the data. In contrast, the random split of 80% for training and 20% for testing may cause the model to be influenced by fluctuations in data distribution. The application of SMOTE improves Recall; however, it may sometimes reduce Accuracy, reflecting the fact that synthetic sample generation enhances the model's ability to identify positive cases while potentially lowering overall predictive accuracy.

The results of Scenario 2, in which the Skin Thickness variable is removed, outperform those of Scenario 1, indicating that excluding Skin Thickness does not significantly affect model performance. In contrast, Scenario 3 demonstrates diminished performance compared with Scenario 1, as Insulin is a critical variable for diabetes diagnosis. Removing such

an important feature limits the model's learning capability and leads to degraded predictive performance.

4. Discussion

It can be observed that the results are strongly influenced by the applied preprocessing strategy. In Scenario 1 (Table 1), all variables are retained and missing values are imputed using the median, resulting in a Recall close to 100% and an Accuracy of approximately 96.61% for the SMOTE_RF model. This indicates that almost no true positive cases are missed, although the risk of false positives may increase (Fig. 7). When Skin Thickness is removed in Scenario 2 (Table 2), Recall decreases slightly to 92.11%, while Accuracy increases to 95.76% with the XGBoost model, suggesting that eliminating a noisy feature enhances model stability without substantially compromising detection capability. In contrast, in Scenario 3 (Table 3), where Insulin is excluded, both Accuracy and Recall decline significantly, thereby confirming the indispensable role of Insulin in diabetes discrimination. Accordingly, Scenario 1 is considered optimal when the primary objective is to minimize missed disease cases, whereas Scenario 2 is more suitable when a balance between high accuracy and reduced false alarms is desired.

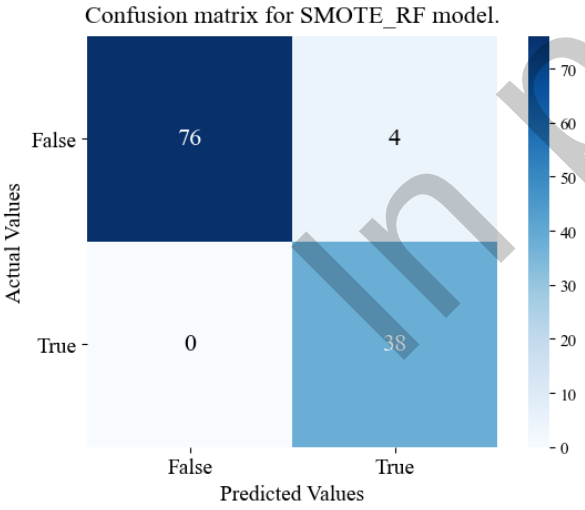

Fig. 7. Confusion matrix

Table 4. Comparison of the results with other studies

Authors	Year	Model	Accuracy
Jobeda Jamal Khanam [22]	2021	Neural Network	88.6%
Hafsa Binte Kibria [23]	2022	RF + XGB	90%
Shimpi, J.K. [24]	2024	KNN + SVM + RF	94.27%
Mahmoud Y. Shams [25]	2025	RFE-GRU	90.05%
Borislava Toleva [26]	2025	SVM	95.5%
Our study		RF XGB	**96.61%** **95.76%**

From Table 4, it can be observed that the PIMA dataset has been extensively utilized by numerous researchers employing various machine learning techniques for the diagnosis and prediction of diabetes. Although performance results vary considerably across studies, our research demonstrates promising outcomes, with the Random Forest classifier achieving the highest accuracy of 96.61%.

Although deep learning approaches such as CNNs [12] and CLSTM models [6] have demonstrated strong predictive performance, ensemble machine learning models offer practical advantages in clinical settings. Tree-based methods such as Random Forest and XGBoost require smaller datasets, are computationally less demanding, and provide greater interpretability through feature importance analysis–an essential requirement for transparent clinical decision-making. Therefore, while deep learning models may achieve competitive performance, the proposed ensemble approach offers a favorable balance between accuracy, interpretability, and computational efficiency. In addition, deploying such predictive models in real-time clinical workflows presents several practical challenges. Unlike static research datasets, real-world clinical environments often involve incomplete, delayed, or dynamically updated patient information. Handling missing values in real-time scenarios may require adaptive imputation strategies or integration with electronic health record systems to ensure data completeness. Moreover, the implementation process must ensure usability for physicians, comply with regulatory requirements, and meet data privacy standards.

Nevertheless, this study still has several limitations. First, the PIMA dataset contains only a limited number of basic biomedical features, such as glucose level, BMI, and number of pregnancies. It does not incorporate other clinically significant risk factors, including lifestyle behaviors (e.g., smoking status, alcohol consumption), dietary patterns, physical activity levels, or detailed family medical history. The absence of these variables may restrict the model's ability to fully capture the multifactorial nature of diabetes risk. Second, missing values were imputed using the median of each variable. Although median imputation is robust to outliers and commonly adopted in clinical datasets, it may reduce data variability and potentially introduce bias. Alternative imputation strategies, including multiple imputation or model-based imputation techniques, could be explored in future research to further enhance model reliability. Third, the proposed models were trained and evaluated exclusively on the PIMA dataset. Since this dataset consists of a specific population group, the results may not be fully generalizable to broader or more diverse populations. Differences in ethnicity, gender distribution, lifestyle, and healthcare conditions across regions may affect model performance. In addition to these data-related limitations, methodological

considerations should also be acknowledged. Although Grid Search was effective for hyperparameter optimization in this study due to the relatively small dataset size, its computational cost increases exponentially with the number of hyperparameters and dataset dimensionality. For large-scale or high-dimensional clinical datasets, more efficient optimization strategies such as Random Search or Bayesian Optimization may provide comparable or superior performance with reduced computational burden.

5. Conclusion

Diabetes mellitus causes numerous serious complications, including life-threatening conditions; therefore, early detection of the disease is critically important. In the context of the rapid expansion of artificial intelligence (AI) in healthcare, machine learning algorithms have increasingly demonstrated their value as effective tools to support clinicians in screening and early diagnosis. In this study, two models–Random Forest and XGBoost–were evaluated on the same dataset. The results indicate that Random Forest outperformed XGBoost, achieving the highest accuracy of 96.61%, while XGBoost also demonstrated strong performance with an accuracy of approximately 95.76%. These findings further highlight the potential role of AI as a valuable decision-support tool in the screening and early detection of diabetes mellitus..

References

[1] Diabetes in Viet Nam, World Health Organization.

[2] T. Jensen and T. Deckert, Diabetic retinopathy, nephropathy and neuropathy. Generalized vascular damage in insulin-dependent diabetic patients., Horm Metab Res Suppl., vol. 26, pp. 68–70, 1992.

[3] P. N. Thotad, G. R. Bharamagoudar, and B. S. Anami, Diabetes disease detection and classification on Indian demographic and health survey data using machine learning methods, Diabetes & Metabolic Syndrome: Clinical Research & Reviews, vol. 17, is. 1, Mar. 2023, Art. No. 102690. https://doi.org/10.1016/j.dsx.2022.102690

[4] M. S. Reza, U. Hafsha, R. Amin, R. Yasmin, and S. Ruhi, Improving SVM performance for type II diabetes prediction with an improved non-linear kernel: Insights from the PIMA dataset, Computer Methods and Programs in Biomedicine Update, vol. 4, 2023, Art. No. 100118. https://doi.org/10.1016/j.cmpbup.2023.100118

[5] Q. Liu, M. Zhang, Y. He, L. Zhang, J. Zoi, Y. Yan, and Y. Guo, Predicting the risk of incident type 2 diabetes mellitus in chinese elderly using machine learning techniques, Journal of Personalized Medicine, vol. 12, iss. 6, p. 905, May 2022, https://doi.org/10.3390/jpm12060905

[6] P. B. K. Chowdary and R. U. Kumar, An effective approach for detecting diabetes using deep learning techniques based on convolutional LSTM networks, International Journal of Advanced Computer Science and Applications (IJACSA), vol. 12, iss. 4, 2021.

[7] U. M. Butt, S. Letchmunan, M. Ali, F. H. Hassan, A. Baqir, and H. H. R. Sherazi, Machine learning based diabetes classification and prediction for healthcare applications, Journal of Healthcare Engineering, vol. 2021, iss. 7, Art. no. 9930985, 2021. https://doi.org/10.1155/2021/9930985

[8] R. Rastogi and M. Bansal, Diabetes prediction model using data mining techniques, Measurement: Sensors, vol. 25, Feb. 2023, Art. no. 100605. https://doi.org/10.1016/j.measen.2022.100605

[9] M. Bukhari, B. F. Alkhamees, S. Hussain, A. Gumaei, A. Assiri, and S. S. Ullah, An Improved artificial neural network model for effective diabetes prediction, Complexity, Hindawi, vol. 2021, pp. 1–10, 2021. https://doi.org/10.1155/2021/5525271

[10] A. Saeed, M. S. Ajmal, M. U. A. Khan, and W. T. Toor, Enhancing early detection of diabetes using machine learning algorithms, in 2024 1st International Conference on Innovative Engineering Sciences and Technological Research (ICIESTR), Muscat, Oman, pp. 1–6, May. 2024. https://doi.org/10.1109/ICIESTR60916.2024.10798412

[11] H. B. Kibria, Md. Nahiduzzaman, Md. O. F. Goni, M. Ahsan, and J. Haider, An Ensemble approach for the prediction of diabetes mellitus using a soft voting classifier with an explainable AI, Sensors, vol. 22, iss. 19, Sep. 2022. https://doi.org/10.3390/s22197268

[12] A. M. M. G. Q. A.-T. S. M. T. H. A. Alawi Alqushaibi Mohd Hilmi Hasan, Type 2 diabetes risk prediction using deep convolutional neural network based-bayesian optimization, Computers, Materials & Continua, vol. 75, no. 2, pp. 3223–3238, Mar. 2023. https://doi.org/10.32604/cmc.2023.035655

[13] S. R. Mishra and S. Dash, Predictive analysis on diabetes detection using pima indian diabetes dataset, INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS, vol. 11, iss. 2, pp. 587–599, Jun. 2024, https://doi.org/10.1109/ICAICT.2011.6110912

[14] Md. K. Hasan, Md. A. Alam, D. Das, E. Hossain, and M. Hasan, Diabetes prediction using ensembling of different machine learning classifiers, IEEE Access, vol. PP, p. 1, Apr. 2020, https://doi.org/10.1109/ACCESS.2020.2989857

[15] M. Heydari, M. Teimouri, Z. Heshmati, and S. M. Alavinia, Comparison of various classification algorithms in the diagnosis of type 2 diabetes in Iran, International Journal of Diabetes in Developing Countries, vol. 36, iss. 2, pp. 167–173, 2016.

https://doi.org/10.1007/s13410-015-0374-4

[16] P. Verma and A. Khatoon, Data Mining Applications in Healthcare: A Comparative analysis of classification techniques for diabetes diagnosis using the pima indian diabetes dataset, in 2024 4th International Conference on Innovative Practices in Technology and Management (ICIPTM), pp. 1–5, 2024. https://doi.org/10.1109/ICIPTM59628.2024.10563296

[17] Blagus, R., Lusa, L. SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics 14, 106 (2013). https://doi.org/10.1186/1471-2105-14-106

[18] M. Bhagat and B. Bakariya, Implementation of Logistic Regression on Diabetic Dataset using Train-Test-Split, K-Fold and Stratified K-Fold Approach, National Academy Science Letters, vol. 45, no. 5, pp. 401–404, 2022, https://doi.org/10.1007/s40009-022-01131-9

[19] Y. Liu, Y. Wang, J. Zhang, New machine learning algorithm: random forest. In Information Computing and Applications (ICICA 2012). Lecture Notes in Computer Science, vol. 7473. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-34062-8_32

[20] A. Tyagi, What is XGBoost algorithm?, Analytics Vidhya. Accessed: Dec. 23, 2025. [Online]. Available: https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/

[21] R. Shah, Tune Hyperparameters with GridSearchCV, Accessed: Dec. 23, 2025. Analytics Vidhya, [Online]. Available: https://www.analyticsvidhya.com/blog/2021/06/tune-hyperparameters-with-gridsearchcv/

[22] J. J. Khanam and S. Y. Foo, A comparison of machine learning algorithms for diabetes prediction, ICT Express, vol. 7, iss. 4, pp. 432–439, Dec. 2021. https://doi.org/10.1016/j.icte.2021.02.004

[23] H. B. Kibria, M. Nahiduzzaman, Md. O. F. Goni, M. Ahsan, and J. Haider, An Ensemble Approach for the Prediction of Diabetes Mellitus Using a Soft Voting Classifier with an Explainable AI, Sensors, vol. 22, iss. 19, Sep. 2022. https://doi.org/10.3390/s22197268

[24] J. K. Shimpi, P. Shanmugam, and S. A. Stonier, Analytical model to predict diabetic patients using an optimized hybrid classifier, Soft comput, vol. 28, iss. 3, pp. 1883–1892, Dec. 2024. https://doi.org/10.1007/s00500-023-09487-w

[25] M. Y. Shams, Z. Tarek, and A. M. Elshewey, A novel RFE-GRU model for diabetes classification using PIMA Indian dataset, Scientific Reports, vol. 15, iss. 1, Jan. 2025, https://doi.org/10.1038/s41598-024-82420-9

[26] B. Toleva, I. Atanasov, I. Ivanov, and V. Hooper, An Effective methodology for diabetes prediction in the case of class imbalance, Bioengineering, vol. 12, iss. 1, Jan. 2025, https://doi.org/10.3390/bioengineering12010035