Golden Sun-Rise
International Journal of Multidisciplinary on Science and Management
ISSN: 3048-5037/ Volume 2 Issue 4 Oct-Dec 2025/ Page No: 84-88
Paper Id: IJMSM-V2I4P110/ Doi:10.71141/30485037/V2I4P110

Research Article

Improvements of Performance of Diabetes Prediction Using Combined Machine Learning Models and Explainable AI Techniques

Pham Tuan Khanh[1], Vu Xuan Manh[2], Hoang Thanh Son[3], Vu Gia Tue[4], Nguyen Hoang Phong[4], Nguyen Minh Huy[5], Tran Ngoc Phuong Linh[6]

[1]Nguyen Hue High School For The Gifted, Ha Noi, Viet Nam.

[2]Hanoi University of Science and Technology, Hanoi, Vietnam.

[3]High School for Gifted Students, Hanoi National University of Education, Hanoi, Vietnam.

[4]HUS High School for Gifted Students, Hanoi, Vietnam.

[5]VNU-HCM High School for the Gifted, Ho Chi Minh City, Vietnam.

[6]Hanoi-Amsterdam Hight School for the Gifted, Hanoi, Vietnam.

Received: 23 October 2025 Revised: 25 October 2025 Accepted: 27 October 2025 Published: 29 October 2025

Abstract - *Diabetes is one of the most dangerous and non-communicable diseases affecting 537 million people worldwide. It affects the pancreas making the body unable to produce insulin (a substance needed to maintain blood glucose level). One of its symptoms is increased urinal level. Early and accurate prediction of diabetes mellitus can help reduce its effect but predicting it is still a challenge for medical doctors. To address this, we apply an AI model using the Pima Indian dataset and data from female patients in Bangladesh to help professionals on gaining preliminary knowledge about the disease on their patients. We also used a semi-supervised technique to fill in missing insulin information of the private dataset. SMOTE and ADASYN were employed to solve the class imbalance problem. The combination of Machine learning classifiers and ensemble techniques were applied to compare the performance of different algorithms to determine which gives the best result. After training, the proposed method obtained an accuracy of 83%. The paper has shown the promising applications in healthcare.*

Keywords - *Diabetes prediction, Healthcare, Machine learning.*

I. INTRODUCTION

Diabetes is a chronic disease resulting from many factors such as high body weight, physical inactivity, and high blood pressure [1]. It attacks the pancreas making the body incapable of producing insulin, causing complications such as increased urine, damaged skin, and kidney failure. In Bangladesh, approximately 7.10 million people had suffered from this disease in 2019; globally, around 537 million people had been diagnosed with diabetes according to 2021 IDF (International Diabetes Federation) data. A substantial body of work has explored diabetes prediction using machine-learning methods. Most studies relied on the Pima Indians Diabetes (PIDD) dataset, with comparatively few using institution-specific data or merged datasets. While many investigations combined multiple algorithms with varied preprocessing pipelines for automated diabetes detection, they commonly emphasized a single accuracy metric, centered exclusively on the PIDD dataset, and gave limited attention to model explainability within their AI frameworks [2]. In this paper, we employ both machine learning and explainable AI to identify diabetes using a private dataset collected from female employees of a local textile industry in Bangladesh alongside the PIDD dataset. Missing entries are imputed with feature-wise means, and data are partitioned using a holdout validation strategy. We evaluate several classifiers Decision Tree, Logistic Regression, K-Nearest Neighbors, Random Forest, Support Vector Machine, and ensemble approaches using precision, recall, and F1-score. The top-performing model is then integrated into a smartphone application. In this paper, the prediction of diabetes mellitus has been made by providing a unique dataset of 203 samples with six variables such as pregnancy, glucose, blood pressure, skin thickness, BMI, age and final outcome of diabetes without disclosing the similarities with the Pima Indian dataset. The problem of class imbalance was addressed using SMOTE and ADASYN. The training was also done by hyperparameter tuning. Another explainable AI method was diabetes prediction, which we used as well. The fusion approach of machine learning models are useful to enhance the performance of suggested method. It is split into sections as follows: Section 2 is a description of the proposed diabetes prediction system; Section 3, the final results are presented; Section 4, conclusion and future improvements are described.

II. PROPOSED METHOD

This part of the text illustrates the procedures and implementations of machine learning techniques to create our automatic diabetes prediction system. Figure 1 shows the different steps of this work.

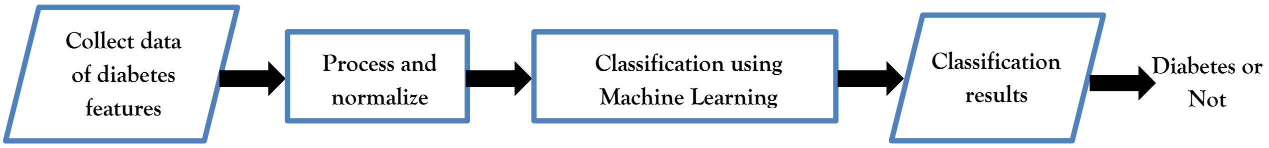

Figure 1. Steps of the Proposed Method

A. Dataset

We evaluate the proposed method on two public datasets: Prima Indian and RTML [3]. Features of the datasets are described as follows:

- Pregnancies: Number of times pregnant
- Glucose: Plasma glucose concentration in an oral glucose tolerance test
- BloodPressure: Diastolic blood pressure (mm Hg)
- SkinThickness: Triceps skin fold thickness (mm)
- Insulin: 2-Hour serum insulin (mu U/ml)
- BMI: Body mass index
- DiabetesPedigreeFunction: Diabetes pedigree function
- Age: Age (years)
- Outcome: Class variable (0 or 1)

B. Data Processing

To combine the dataset, we found some exceptional zero values (skin thickness and BMI). These values were replaced by the mean value.

C. Machine Learning Classifiers

In this study, a broad set of machine-learning and ensemble approaches were utilized: the GridSearchCV tuning framework, Decision Tree, K-Nearest Neighbors (KNN), Random Forest, Support Vector Machine (SVM), Logistic Regression, AdaBoost, XGBoost, Voting Classifier, and Bagging [4]. GridSearchCV is a scikit-learn procedure for hyperparameter optimization that exhaustively tests combinations from a predefined parameter grid using cross-validation. It streamlines model selection by fitting and evaluating each configuration against a chosen metric (e.g., accuracy, F1-score) and returning the best-performing setup [4]. A Decision Tree is a supervised model that partitions the feature space through a series of hierarchical splits. Internal nodes encode rules based on feature thresholds, while terminal leaves represent predicted classes (for classification) or continuous outputs (for regression) [5] [6].

The KNN Classifier is a non-parametric, instance-based method that labels a query point according to the majority class among its k closest training samples, where closeness is typically computed via a distance measure such as the Euclidean metric [5] [7]. Random Forest is an ensemble technique of learning that builds a series of decision trees in the course of training, and combines their predictions (majority vote in the classification, average in the regression). It brings the element of randomness through bootstrap sampling and random feature selection [6] [8].

A Support Vector Machine is an algorithm to detect the optimal hyper plane to partition the data points of divergent classes with the most significant margin. It applies the so-called kernel trick to data in the higher dimensional space in situations where the data cannot be separated in a linear manner [9]. Logistic Regression is a linear regression that can be applied in binary classification (equivalent to multiclass with softmax). It estimates the likelihood of a datum point to be a member of a given set of classes based on the use of logistic (sigmoid) function [10]. AdaBoost is an ensemble learning algorithm, a collection of weak classifiers (usually decision stumps) used together to form a strong one. It puts more emphasis on the misclassified data points in each iteration to concentrate on the challenging cases [11].

XGBoost is an optimized gradient boosting model which takes the form of an ensemble of decision trees built sequentially, minimizing a loss function via gradient descent. It involves regularization so as to avoid overfitting and parallel processing to ensure efficiency [12]. Voting Classifier Voting Classifier A voting Classifier is an ensemble technique, and it uses a combination of multiple predictions made by classifiers, with either majority voting (hard voting) or averaging predicted probabilities (soft voting) to get the final prediction [13]. Bagging is a form of ensemble method that uses several instances of a underlying base learner (e.g., decision trees) and uses random subsets of the training data (with replacement) to train the instances and (via voting or averaging) [14].

D. Explainable AI (XAI)

Explainable AI (XAI) refers to methods and techniques in artificial intelligence (AI) and machine learning (ML) that make the decisions or predictions of AI models transparent, interpretable, and understandable to humans. In contrast to classical black-box models (e.g. deep neural networks), that make precise, yet opaque predictions, XAI is designed to give an insight into how and why a model makes certain prediction. This is important in establishing trust, accountability and regulatory provisions in areas such as healthcare, finance and legal systems [15].

III. PERFORMANCE EVALUATION OF THE PROPOSED METHOD FOR DIABETES PREDICTION

This section presents the outcomes of our automated diabetes prediction framework. We assess multiple ML classifiers using the precision metric. Precision is widely adopted to gauge how reliably a model's positive predictions correspond to true positives. Table 1 shows the performance metrics of all classifiers.

Table 1. Performance Comparison of Classifiers for Diabetes Prediction

Method	Precision
kNN	74%
SVM	75%
Voting	77%
Decision tree	78%
XGBoost	81%
Combined models	83%

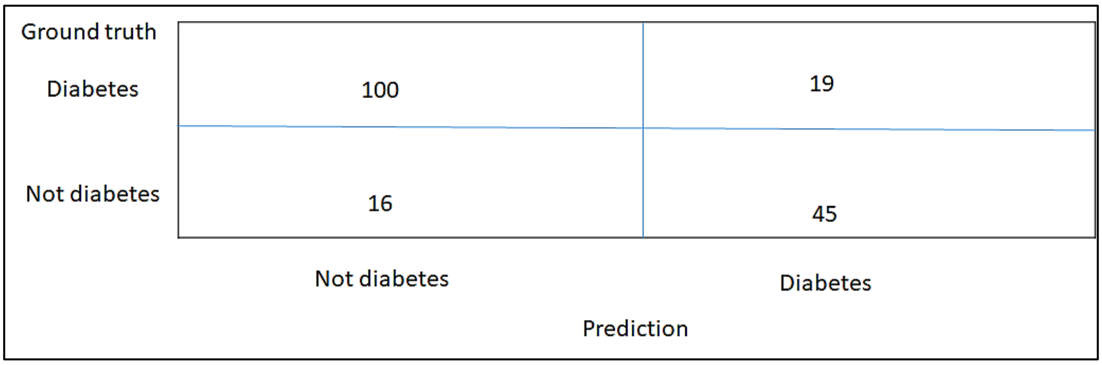

Figure 2. Confusion Matrix of the Classification

Figure 2 depicts the confusion matrix for the XGBoost classifier trained with ADASYN oversampling, while Figure 3 presents its Receiver Operating Characteristic (ROC) curve. The ROC curve summarizes the trade-off between true-positive and false-positive rates and reflects the classifier's discriminative power

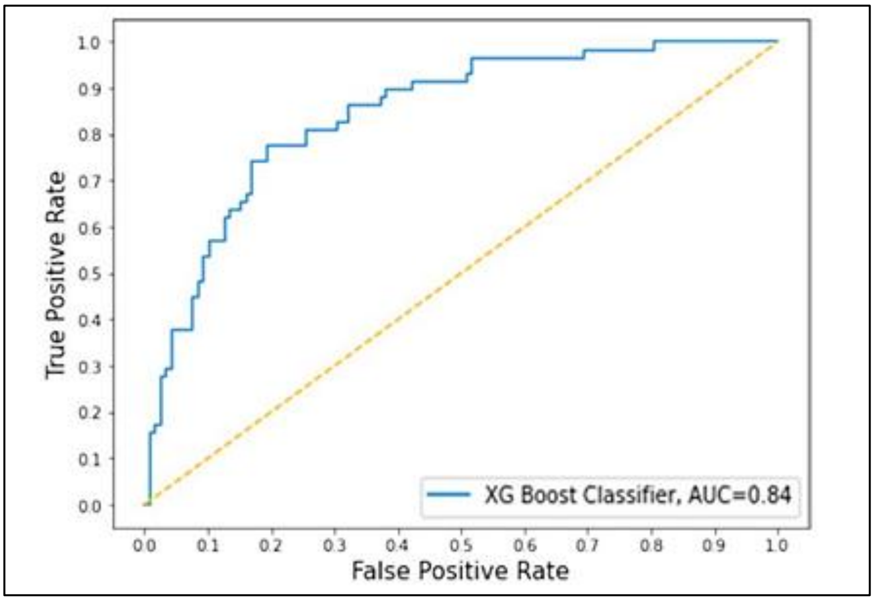

Figure 3. ROC curve and AUC value for XGBoost with ADASYN

According to these figures, the model predicts diabetes correctly with 83% confidence. The model classified this class as a person that has a glucose level of more than 140.25 and involves pregnancies of more than 6. According to obtained results, we can see that the Glucose, BMI and Age features significantly affect diabetes people. Explainable AI techniques with SHAP and LIME frameworks are implemented with figure 4 showing the feature importance.

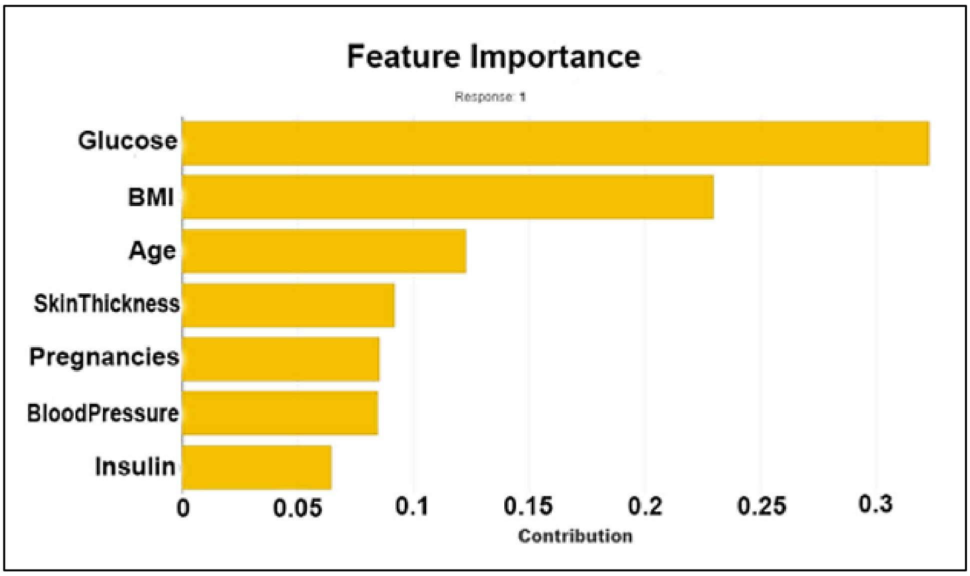

Figure 4. Explainable AI Showing the Importance of Features [16]

IV. CONCLUSION

The paper presents the diabetes classification using machine learning and explainable AI. We have successfully achieved a model using the Pima Indian and RTML dataset with a higher F1 score than previous works. There are some future changes for this work, for example, we recommend getting additional private data with a larger cohort of patients to get better results. The combination of machine learning models helps to improve the performance. Another extension of this work is combining machine learning models with fuzzy logic techniques and applying optimization approaches.

Conflicts of Interest

The authors declare that there is no conflict of interest concerning the publishing of this paper.

V. REFERENCES

1. A. Rghioui, J. Lloret, S. Sendra, and A. Oumnad, "A Smart Architecture for Diabetic Patient Monitoring Using Machine Learning Algorithms," *Healthcare (Switzerland)*, vol. 8, no. 3, 2020. Google Scholar | Publisher Link
2. M.K. Hasan, M.A. Alam, D. Das, E. Hossain, and M. Hasan, "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers," *IEEE Access*, vol. 8, pp. 76516–76531, 2020. Google Scholar | Publisher Link
3. "Diabetes Data Set," *Kaggle*. Online: https://www.kaggle.com/datasets/mathchi/diabetes-data-set.
4. S. Raschka and V. Mirjalili, *Python Machine Learning* (4th ed.), Packt Publishing, 2022. Google Scholar | Publisher Link
5. J. Bergstra and Y. Bengio, "Random Search for Hyper-Parameter Optimization," *Journal of Machine Learning Research*, vol. 13, pp. 281–305, 2012. Google Scholar | Publisher Link
6. L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone, *Classification and Regression Trees*, Wadsworth, 1984. Google Scholar | Publisher Link
7. T.M. Cover and P.E. Hart, "Nearest Neighbor Pattern Classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967. Google Scholar | Publisher Link
8. L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. Google Scholar | Publisher Link
9. C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995. Google Scholar | Publisher Link
10. D.W. Hosmer, S. Lemeshow, and R.X. Sturdivant, *Applied Logistic Regression* (3rd ed.), Wiley, 2013. Google Scholar | Publisher Link
11. Y. Freund and R.E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997. Google Scholar | Publisher Link
12. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016. Google Scholar | Publisher Link
13. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning* (2nd ed.), Springer, 2009. Google Scholar | Publisher Link
14. L. Breiman, "Bagging Predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996. Google Scholar | Publisher Link
15. M.T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?' Explaining the Predictions of Any Classifier," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016. Google Scholar | Publisher Link
16. I. Tasin, T.U. Nabil, S. Islam, and R. Khan, "Diabetes Prediction Using Machine Learning and Explainable AI Techniques," *Healthcare Technology Letters*, 2022. Google Scholar | Publisher Link