Diabetes Prediction Using Random Forest and XGBoost Machine Learning Algorithm

Ramesh Prasad Bhatta[1*]
[1] Central Department of CSIT, Far Western University, Nepal
[*]Corresponding Author, Email: rpb.mcs@gmail.com

Abstract

Diabetes mellitus is a prevalent chronic disease with serious global health implications, where timely identification is crucial for effective management and intervention. Accurate prediction of diabetes can greatly enhance patient care by enabling prompt medical responses. In recent years, machine learning techniques have gained attention in the healthcare domain for disease prediction and prognosis. This study investigates the application of Random Forest (RF) and XGBoost (XGB) classifiers for predicting diabetes using the PIMA Indian Diabetes dataset. Data preprocessing methods—including missing value imputation, normalization, feature selection, and upsampling were applied to improve data quality and model accuracy. Hyperparameter tuning was also conducted to further optimize model performance. To enhance predictive capability, a soft voting ensemble integrating RF and XGB was developed, achieving outstanding results with an AUC of 0.91, an accuracy of 0.84, a precision of 0.80, and a recall of 0.92, indicating both strong predictive ability and reliability. The SHAP (Shapley Additive Explanations) value analysis revealed that glucose, age, and BMI were the most influential factors contributing to diabetes risk. The results highlight the potential of ensemble learning methods in healthcare analytics. this study contributes to leverage interpretable machine learning for early disease detection and informed clinical decision-making.

Keywords: *Diabetes, Machine Learning, Prediction, Random Forest, XG Boost Classifier*

Introduction

Diabetes is a metabolic condition defined by elevated blood sugar levels. It can cause serious problems for the heart, kidneys, blood vessels, nerves, and eyes. Globally, around 422 million people are affected by diabetes, with the majority residing in low- and middle-income countries. The disease directly contributes to over 1.6 million deaths annually, highlighting its significant public health burden. Diabetes is considered as major healthcare issue that is affecting the world at a rapid and alarming rate (Chandra Sen, P., et al. 2020). "Diabetes mellitus is deadliest and is caused by a set of metabolic disorders that occur when the body cannot produce any or enough insulin or cannot effectively use the insulin it produces". When a person has abnormally high blood glucose levels because of either inadequate insulin manufacturing or an inappropriate cell response to insulin, they have diabetes mellitus, one of the metabolic illnesses. Diabetes is recognized as a metabolic illness that arises when blood sugar levels are elevated for an extended period of time. Glucose is our body's most essential energy source since it aids in the development of our muscles, tissues, and cells. Upon leaving the glucose unabsorbed in the body, Diabetes develops as a result of elevated blood sugar levels in cells are unable to use it.

Types of Diabetes

1. **Type 1-Diabetes mellitus (T1DM):** It is the most common kind of diabetes and is identified by the body's inadequate production of insulin. The illness can strike at any age, although children and teenagers are the ones who get it most often (Sarwar, M. A., et al. 2018). In this type of diabetes, pancreas will produce insulin that helps human organs to energize through sugar level in the blood cells. But there may be a chance that pancreas might be producing little amount of insulin or no insulin. Insulin injections are commonly used for controlling Type-1 diabetes. Type-1 diabetes is common in any aged people but it most affects the people among under age 30. Particularly if the patient's heredity having Type-1 diabetes will lead to higher risk. Statistically below 10% of the people impacted by this particular form of diabetes.

2. **Type 2-Diabetes mellitus (T2DM)**: It is the most prevalent kind of diabetes and is distinguished by the body's insufficient synthesis of insulin. All age groups are affected, and patients frequently show signs of obesity, overweight, urination, etc., which are associated with the Insulin resistance (Mujumdar, A., & Vaidehi, V. 2019). Historically, the adults are most commonly affected by Type 2

diabetes. Statistics revealed that betwixt 90-95 percent of the people will be affected by type 2 diabetes. Diet through weight management and exercise are the common ways to control Type 2 diabetes. Yet, medications or injections may be considered as remedy for lowering the glucose level

3. **Gestational diabetes mellitus (GDM):** The kind of diabetes that causes pregnant women to have hyperglycemia. This kind of diabetes raises the mother's and the fetus's risk of developing type-2 diabetes. In general, the pregnant women will have Gestational diabetes who never had a diabetes in their lifetime. The glucose level will be high when the women get pregnancy. The baby has higher glucose level at the time of pregnancy. Changes in hormone will also leads to high glucose level in blood that affects the action of insulin.

4. **Prediabetes (PD):** Genetic abnormalities that result in increased insulin production are the cause of this form of diabetes., side effects of chemicals, or increase in other hormonal levels in the body. Lifestyle habits and demographic factors are examined and reported the main indicators that play an important role to control and manage Type-2 Diabetes Mellitus. Diet and exercise play an important role to avoid or manage the T2DM, it can reduce the complications of even those people who are at high risk of being involved towards diseaseIoT is a paradigm where the interconnection of objects with myriad sensing and actuating devices through the Internet provides the ability to collect, share, and analyze information for enabling innovative applications (Gubbi, Buyya, Marusic, & Palaniswami, 2013). By enabling interaction with a wide variety of objects such as, appliances, surveillance cameras, wearables, smart phones, industrial sensors, and vehicles, IoT can facilitate the development of many new services for citizens, businesses, and governments. This paradigm finds application in multiple domains, such as transportation, energy and utilities, education, healthcare, physical infrastructure, public safety and defense, among others.

Problem Statement

Machine learning has become an increasingly valuable tool in the field of medical diagnosis, offering promising results in the prediction of various health conditions. In the context of diabetes prediction, the PIMA Indians Diabetes Dataset has been widely used as a standard benchmark for developing and comparing predictive models. Researchers have applied different algorithms such as logistic regression, support vector machines, and neural networks, with ensemble methods often standing out for their higher robustness and predictive accuracy.

However, despite the considerable progress made in this area, several important challenges remain unaddressed. Many existing studies place heavy emphasis on improving accuracy, often overlooking the equally important aspects of model interpretability and clinical relevance. Additionally, issues such as missing data and class imbalance within the dataset are not consistently handled, which can significantly affect the reliability of model outcomes.

Research Gaps

1. Limited interpretability and clinical use: Most studies on diabetes prediction with the PIMA dataset focus mainly on accuracy, giving little attention to how predictions are made or how the results can support real clinical decisions.
2. Inconsistent data handling and weak model optimization: Existing research often treats missing data and class imbalance inconsistently, and many ensemble models are not well-optimized or thoroughly interpreted.

Research Questions

1. How does an ensemble of Random Forest and XGBoost enhance the accuracy of diabetes prediction?
2. How can SHAP-based interpretability analysis make machine learning models more useful for healthcare applications?

To address these gaps, this research focuses on two main objectives:

Objectives of the study

- Developing an ensemble model that integrates the strengths of Random Forest and XGBoost to enhance predictive performance.
- Implementing an interpretability approach using SHAP analysis combined with clinical risk stratification.

Literature Review

In this approach, they suggest diabetic patients and analyzing them by using several diabetes characteristics to forecast the onset of diabetes. An analysis-based intelligent diabetes disease prediction system of diabetes using a database of diabetes patients. (Shetty et al. 2017). This developed a system that uses a Random Forest algorithm for diabetes prediction to do early diabetes prediction for a patient and found the improved accuracy. The results showed that the prediction system can accurately, rapidly, and most importantly, predict the onset of diabetes. The recommended model yields the best results for diabetic prediction. (VijiyaKumar et al.2019). (Hasan et al.2020) proposed a comprehensive data preprocessing framework for diabetes prediction, which integrates outlier detection, feature selection, and missing value imputation to improve data quality. The study employed a variety of machine learning classifiers and multi-layer perceptrons, to develop predictive models for diabetes. In addition, the authors introduced an ensemble classifier that combines the outputs of multiple models through a weighted voting mechanism, enhancing prediction accuracy.

This study proposed a machine learning approach for diabetes prediction that combines the XGBoost hybrid fusion method with the Random Forest (RF) algorithm. The study emphasized the role of feature selection in enhancing the performance of the fusion model by identifying the most relevant factors for diabetes prediction. The authors demonstrated that optimizing the model through feature selection could improve patient outcomes and support personalized diabetes management (Gonzalez et al.2021). (Kumari et al. 2021) The proposed method was calculated experimentally using the state-of-the-art. Logistic regression, Random Forest, SVM, Naïve Bayes, AdaBoost, GB, cat boost, XGB and bagging are examples of base classifiers that serve the same function (Kumari et al. 2021). Based on the PIMA diabetes dataset, the results demonstrate that the proposed collective methodology could produce precision, accuracy, F1-score, and recall outcomes with 73.48%, 79.04%, 80.6%, and 71.45%, respectively.

The hybrid approach for diabetes diagnosis that integrates using questionnaire-based data. The system employs an MLR-RF method for feature selection and XGBoost for classification. The dataset comprises hospital records of 520 individuals in Sylhet, Bangladesh, including 200 control cases and 320 diabetes cases. The proposed method demonstrated high reliability and efficiency for diabetes prediction, achieving an accuracy of 99.2%, an AUC of 99.3%, and a prediction time of 0.04825 seconds (Gundogdu 2023). (Mujumdara and Vaidehi 2019) applied multiple machine learning algorithms to the dataset and compared their performance. Among the classifiers, Logistic Regression achieved the highest accuracy of 96%, while AdaBoost proved to be the most effective model during pipeline implementation, reaching an accuracy of 98.8%. The study highlights the comparative evaluation of different machine learning techniques for diabetes prediction and demonstrates the potential of ensemble methods like AdaBoost in improving predictive performance. This highlights the importance of exploring advanced machine learning techniques for accurate diabetes prediction (Laxmikant K et al., 2023). By leveraging the capabilities of XGBoost, researchers can improve the performance of predictive models, enabling early detection and personalized healthcare management for individuals at risk of diabetes. (Chou, Hsu, and Chou 2023) examined 15,000 women aged 20 to 80 using outpatient examination data collected from a Taipei Municipal medical center between 2018–2020 and 2021–2022.The study analyzed eight predictor variables of PIMA dataset for diabetes prediction. The results indicated that the two-class boosted decision tree achieved the highest predictive performance, with an AUC of 0.991, outperforming the other models, whose top comparative AUC was 0.976. This study investigates method for diabetes prediction using health-related indicators from the dataset. The results showed that Naïve Bayes achieved higher predictive performance, with an average accuracy of 76.07%, precision of 73.37%, and recall of 71.37% (Febrian M. E. et al. 2023). This study aimed to assess the Random Forest and XGBoost algorithms' performance in diabetes classification. (N. J. Dzira et al., 2025). The findings showed that feature selection considerably improved both models' accuracy, with GA outperforming PSO. In particular, PSO increased accuracy by 5.7% to 7.6% and GA increased accuracy by 6.1% to 8.9%. This study established a framework for early warning system and it has capacity to manage diverse medical data. A prediction accuracy of 94.7% was attained by the suggested architecture (Xiong, et al 2025). In order to study diabetes prediction, (Ahmed, Khan, et al. 2025) used the Pima Indians Diabetes Dataset (PIDD) from the Kaggle repository. Four machine learning algorithms were used after exploratory data analysis (EDA) using PCA, heatmaps, and scatter plots. The results showed that 80% accuracy, 82% precision, 88% sensitivity, and 20% of error rate. The results indicated that RF performed the best. According to the study, ML-based prediction models can help identify diabetes early on and stop the disease from getting worse.

Research Methodology

Research involved in this study is aimed at developing a framework for predicting diabetes at the initial phase by making use of classification models. The figure below depicts the study's working plan.

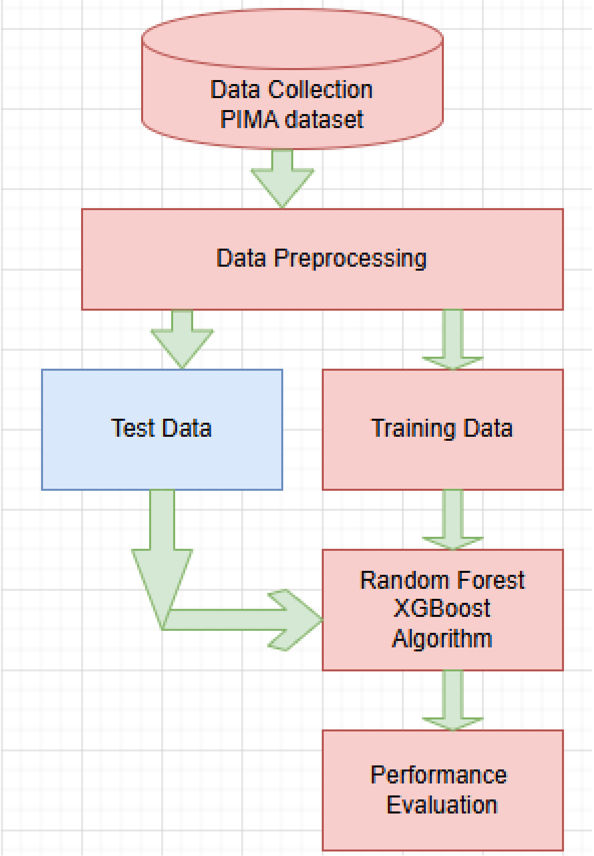

Figure 1: Working of ML Process in Diabetes Prediction

The following steps comprise the machine learning process:

Data Collection

This study uses the Pima Indian dataset which is publicly accessible on the Kaggle. The dataset contains 768 individual data in the sample range in age from 21 to 81 years. The 268 in total, comes from people who have been diagnosed with diabetes. Eight variables make up the dataset: age, body time index, triceps skinfold thickness, glucose concentration, diastolic blood pressure, number of pregnancies, 2-hour insulin, and history of hereditary diseases.

Data Preprocessing

In machine learning, data preprocessing is the process of converting unstructured data into a format that is appropriate for training and assessing machine learning models. This important stage deals with frequent problems that can greatly affect model performance in real-world datasets, like missing values, inconsistencies, outliers, and irrelevant features. Data Preprocessing involves following activities.

1. Handling Missing Values:

Invalid values (e.g., zero for BMI, glucose, or insulin) were replaced with mean/median imputation.

2. *Handling outliers*:

Extreme values that deviate from the norm and are not consistent with the rest of the data may be present in a dataset. It entails figuring out the IQR, or the difference between the first (Q1) and third (Q3) quartiles. Data points that fall below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR are then classified as outliers.

3.*Standardization/ Normalization*:

Normalization in machine learning is a data scaling technique where numerical features are transformed to fit within a specific range — usually [0, 1] or [-1, 1].

4.*Feature Extraction*

The attributes were standardized to make sure that all attributes were on the same scale in order to enhance the performance of ML models. And hence improves accuracy and performance are enhanced by feature normalization or standardization.

5.*Splitting Datasets*

To train the ML models, the dataset was divided into a Training Set (70%) of the total. Model evaluation on unseen data was conducted using the Testing Set (30%).

6.*ML Models Development*

After splitting the dataset, the model was deployed on Python.

Machine Learning algorithms

In order to train the model and predict diabetes, this study used Random Forest and XGBoost machine learning models.

Random Forest

Random Forest is a machine learning technique that uses a lot of decision trees to produce better predictions. Regression averaging or classification voting are used to aggregate the results of each tree's analysis of discrete random data segments technique that enhances prediction robustness and accuracy. The Random Forest algorithm is an ensemble approach that increases prediction accuracy by utilizing a large number of decision trees. A subset of the data is used to create each tree, and the outcome is decided by a majority vote from all the trees. This method lessens the possibility of overfitting, which frequently happens with the single decision tree approach.

Random Forest works as follows:

1. Random Sampling:

Data samples are randomly selected with replacement (bootstrap) to form training subsets equal in size to the original dataset.

2. Random Feature Selection:

A small subset of features is randomly chosen from the available set for tree construction.

3. Decision Tree Construction:

A decision tree is built using algorithms like ID3, C4.5, or CART by splitting data on the most informative features.

4. Ensemble Formation:

Steps 1–3 are repeated multiple times to create an ensemble of decision trees.

5. Final Prediction:

The overall output is determined by majority voting (classification) or averaging (regression) across all trees.

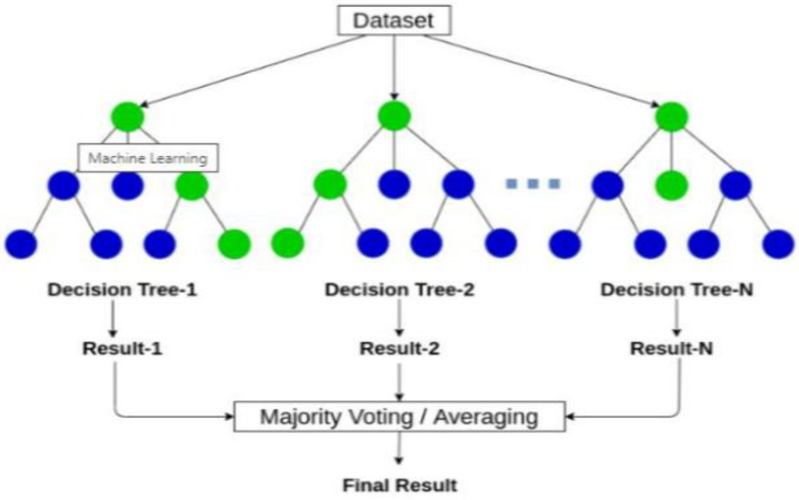

Figure 2: Random Forest Algorithm Diagram

XGBoost Algorithm

XGBoost is an ensemble boosting algorithm that builds a strong predictive model by integrating multiple weak learners. The most powerful ensemble technique for classification and prediction is called XG Boosting. It creates powerful learning models for prediction by combining week learners. The Decision Tree model is used.The performance of the gradient boosting model improves with each iteration.

Algorithm for XGBoost

1. Consider a sample of target values as P.
2. Estimate the error in target values.
3. Update and adjust the weights to reduce error M.
4. P[x] =p[x] +alpha M[x]
5. Model Learners are analyzed and calculated by loss function F
6. Repeat steps till desired & target result P.

Shapley Additive Explanations

Shapley Additive explanations (SHAPs) are a useful technique for interpretability (Lundberg, S. M., & Lee, S. I. 2017). Cooperative game theory is the source of the Shapley values, which are used to determine the contribution of each player (or feature) to the overall game (or prediction model). Understanding how each dataset component influences the prediction outcome is essential for identifying which elements are most predictive of diabetes. The SHAP value for each feature is calculated using the formula below:

$$\phi j = \sum_{S \subseteq N \setminus \{j\}} = \frac{|S|!(|N| - |S| - 1)!}{|N|!} [\ fx(S \cup \{j\}) - fx(S)] \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 1$$

In the above equation,

- ϕj is the SHAP value for feature j;
- S is a subset of features excluding j;
- N is the total set of features;
- |S| is the number of features in subset S;
- |N| is the total number of features;
- fx(S) is the output for the model using features in S;
- fx(S ∪ {j}) is the model output using features in S along with feature j.

The difference [fx(S ∪ {j})− fx(S)] quantifies the marginal contribution of feature j to the prediction when added to subset S.

Confusion Metrics

A confusion matrix is a matrix that summarizes the performance of a machine learning model on a set of test data. It is often used to measure the performance of classification models, which aim to predict a categorical label for each input instance. True Positive (TP): Correctly predicted positive cases. False Positive (FP): Incorrectly predicted positive cases (actually negative). True Negative (TN): Correctly predicted negative cases. False Negative (FN): Incorrectly predicted negative cases (actually positive).

Performance metrics

Accuracy (ACC):

Accuracy is computed as the number of all correct predictions divided by the total number of the dataset, which is the number of patients that are identified correctly in total in our case.

Accuracy =((TP+TN))/((TN+FP+FN)) ………… 2

Precision:

Precision is computed as the number of correct positive predictions divided by the total number of positive predictions.

Precision=TP/(TP+FP) …………. 3

Recall(Sensitivity)

Recall is computed as the number of correct positive predictions divided by the total number of positives. it is also called Sensitivity or true positive rate (TPR).

Recall=TP/(TP+FN) ……. 4

F1 score:

The F1-score is the harmonic mean of precision and recall, providing a balance between the two metrics. It is calculated as twice the product of precision and recall divided by the sum of precision and recall. it provides the quality of prediction.

F1 score = 2*(Precision*Recall)/(Precision +Recall) ……. 5

Result and Discussion

Dataset Statistics

Two machine learning algorithms were utilized in this research; the following table lists the statistical values for the different attributes in the dataset.

Table 1: Statistical summary of different variables in dataset (from Python)

	Count	Mean	Std	Min	25%	50%	75%	Max
Pregnancies	768	3.84	3.37	0.00	1.00	3.00	6.00	17.00
Glucose	768	120.89	31.97	0.00	99.00	117.00	140.25	199.00
Blood Pressure	768	69.11	19.36	0.00	62.00	72.00	80.00	122.00
Skin Thickness	768	20.54	15.95	0.00	0.00	23.00	32.00	99.00
Insulin	768	79.80	115.24	0.00	0.00	30.50	127.25	846.00
BMI	768	31.99	7.88	0.00	27.30	32.00	36.60	67.10
Diabetes Pedigree Function	768	0.47	0.33	0.08	0.24	0.37	0.62	2.42
Age	768	33.24	11.76	21.00	24.00	29.00	41.00	81.00

Table 1 above uses basic quantitative analysis to determine the mean, median, etc. It is counterintuitive in reality that many of those basic statistics have a minimum value of 0. For example, people cannot have blood pressure that is 0. The data's zero values could be the missing values. Following analysis, the following invalid zero values are reported: BMI, insulin, skin thickness, blood pressure, and glucose.

A total of almost three hundred rows would be removed if we choose to eliminate all observations with invalid zero values. If so many rows are dropped, the sample will be too small to carry out training. It is preferable to substitute appropriate values for faulty zero values The mean of the current values is used to replace the missing values of blood pressure, glucose, and skin thickness, and the median of the current values in the corresponding column is used to replace the missing values of insulin, BMI, and skin thickness, based on the distribution and realistic meaning of the five variables.

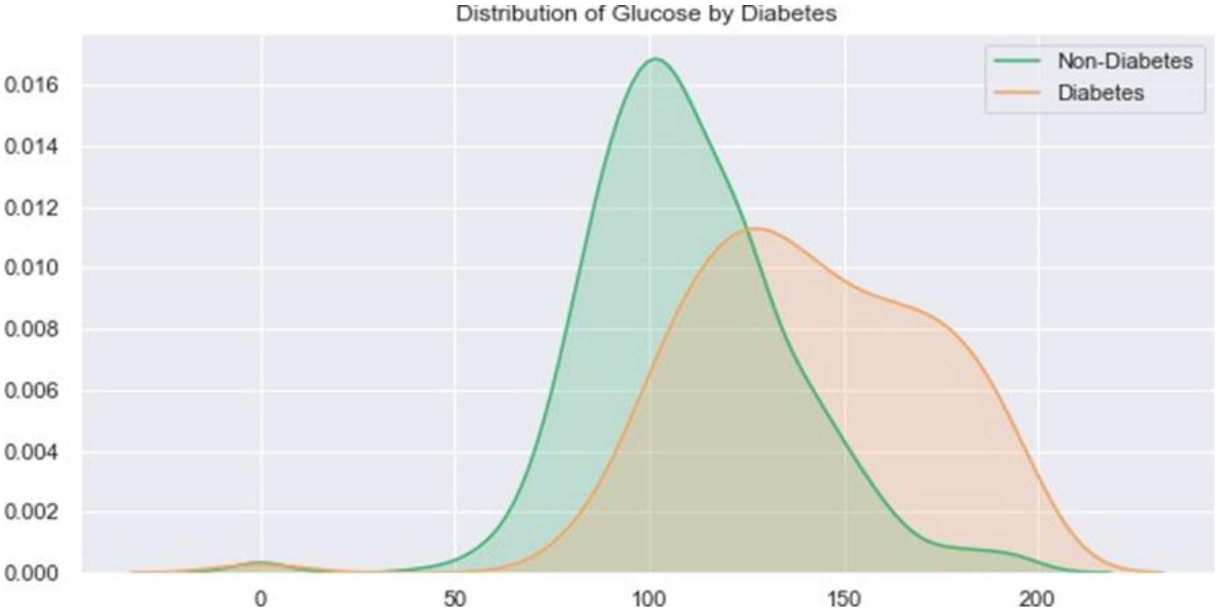

Figure 3(a) Distribution of Glucose by Diabetes

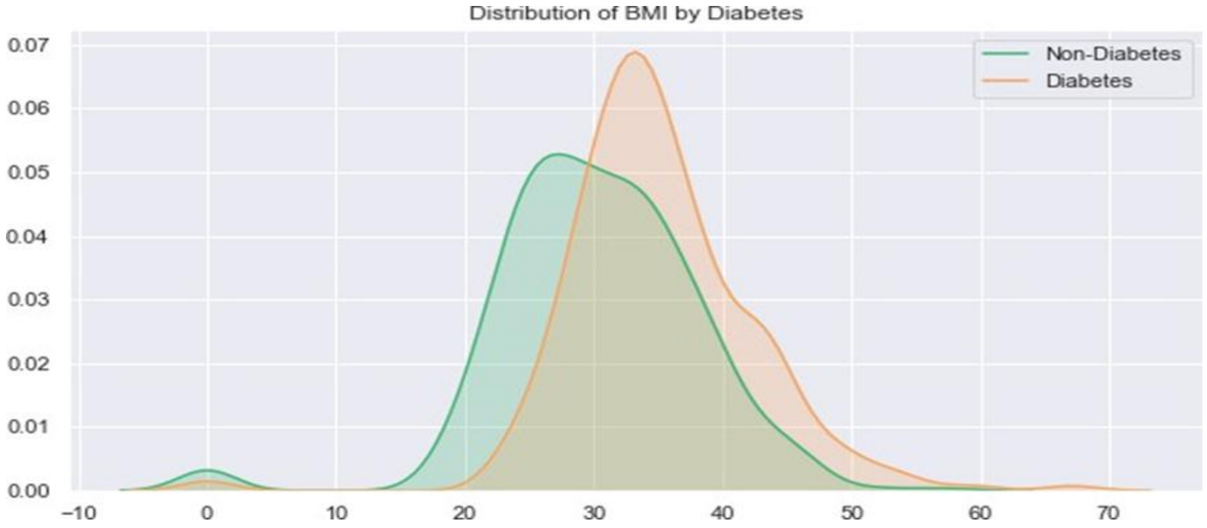

Figure 3(b) Distribution of BMI by Diabetes

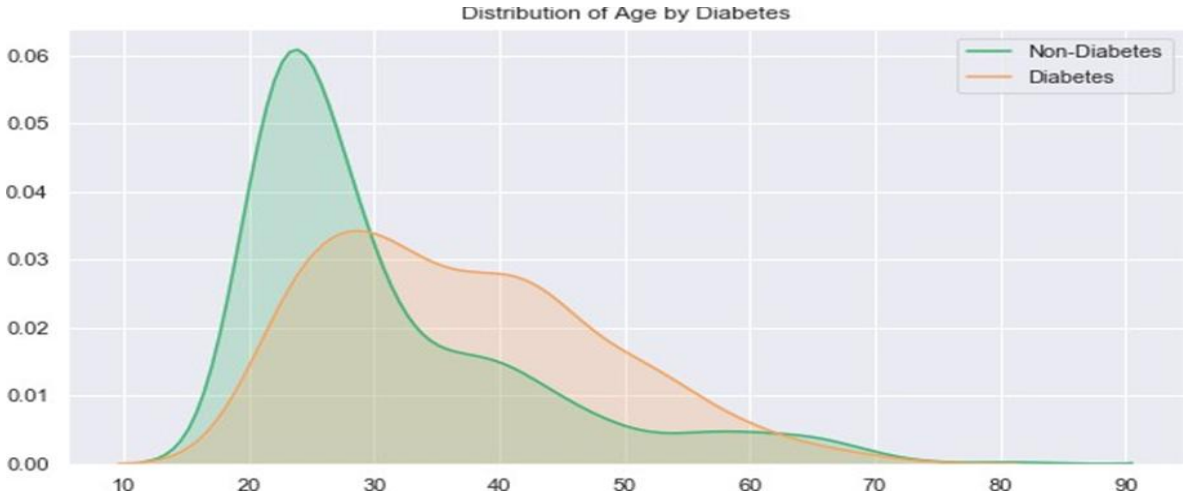

Figure 3(c) Distribution of Age by Diabetes

There is clear separation between the two groups. This plot clearly stated that higher glucose levels strongly correlate with diabetes, making it one of the most important features in prediction models. Histogram of data features in PIMA dataset is illustrated in following diagram.

The figure 4 showed that the majority of glucose readings are concentrated between 80 and 140, according to the histogram. Some patients had significantly higher glucose levels (outliers above 150–200), indicating that the distribution is roughly bell-shaped but slightly tilted to the right. This distribution implies that although most individuals are in the normal-to-borderline range, a sizable portion have high glucose, which may be a symptom of diabetes risk, as glucose is a crucial diagnostic factor for diabetes.

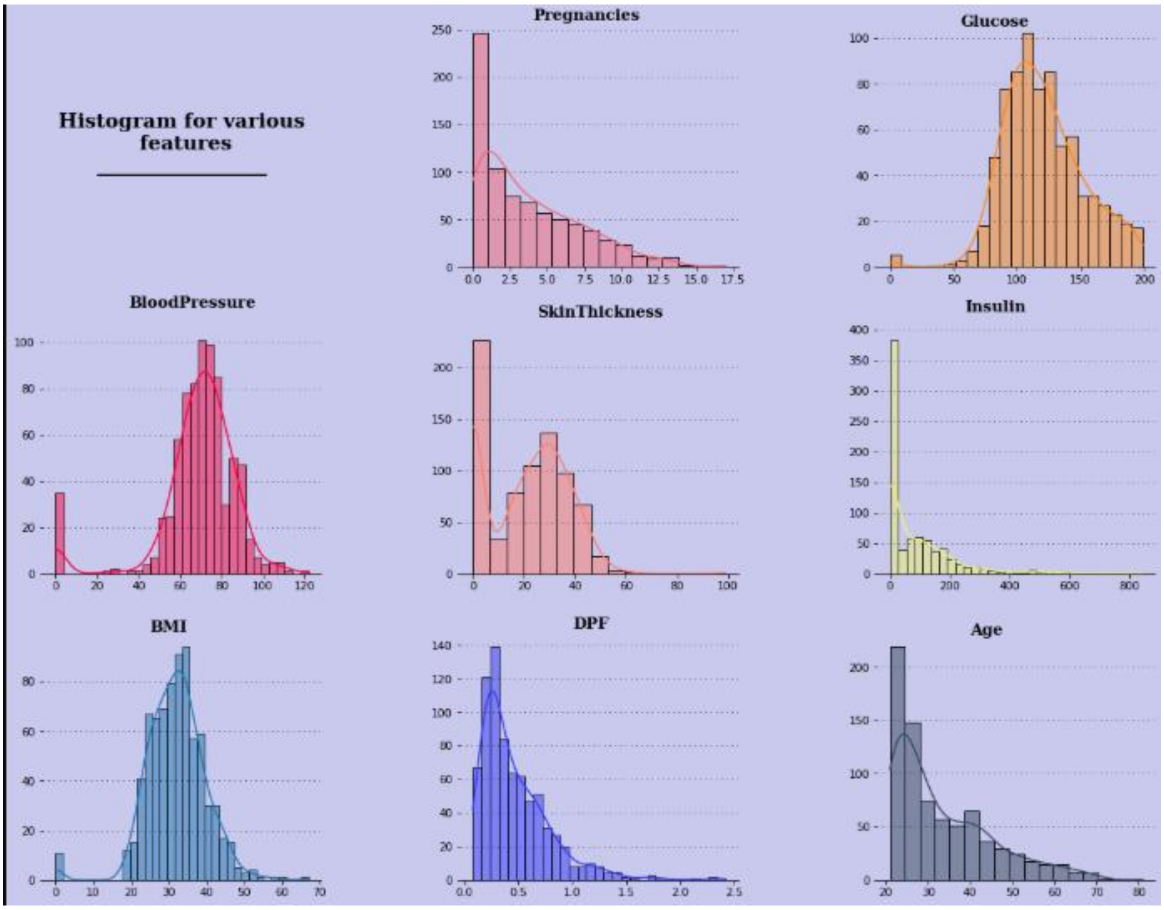

Figure 4 Histogram for PIMA dataset features

Correlation between various features in dataset and most crucial features for diabetic risk are also clearly depicted in following heatmap in figure 5.

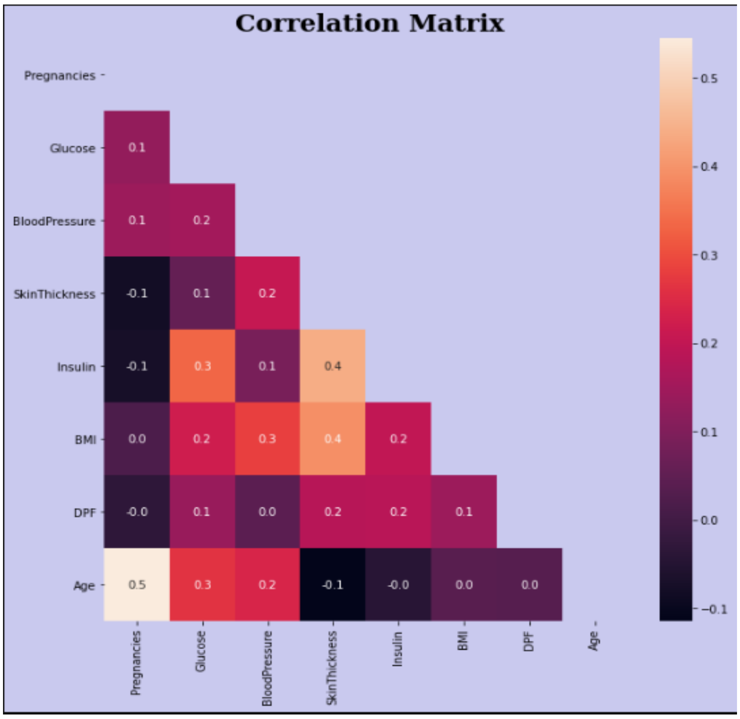

Figure 5 Correlation matrix for features

The diagram confirms that high Glucose levels, high BMI, and older Age are the strongest predictors for a diabetes prediction in this PIMA dataset. Additionally, there is a strong positive link between age and BMI and prognosis. Age and a higher body mass index (BMI) are recognized risk factors for diabetes. The correlation study showed that characteristics skin thickness and blood pressure seem to have very little relation for diabetic.

Confusion Metrics Analysis

Random Forest

The maximum tree depth for random forest training is likewise set at four. The woodland contains one hundred trees. Additionally, computed and displayed in Figure 6 are the confusion matrix and receiver operating characteristics (ROC). The results obtained from the confusion matrix.

Accuracy =(TP +TN)/(TP + TN + FP +FN)

=(39+114)/(39+114+11+28)=79.7%

The precision= (TP /)(TP + FP)=39/(39+11)=78.0%

The recall = (TP)/ (TP + FN)

=39/(39+28)=58.2%

True Negative Rate = $TN/(TN+FN)$

=11/(41+14)+28=80.3%

False Positive Rate (FPR) = $(1-TNR)$ =1−74.5%=19.7%

When it comes to identifying non-diabetes (high TN), the model is excellent. Some cases of diabetes are missed, though (28 FN). It predicts diabetes rather well, albeit it somewhat benefits the class without diabetes. The model's great discriminative power is indicated by the curve's approach to the top-left corner. Visually, it appears to be between 0.85 and 0.90, indicating that the model can effectively differentiate between instances with and without diabetes.

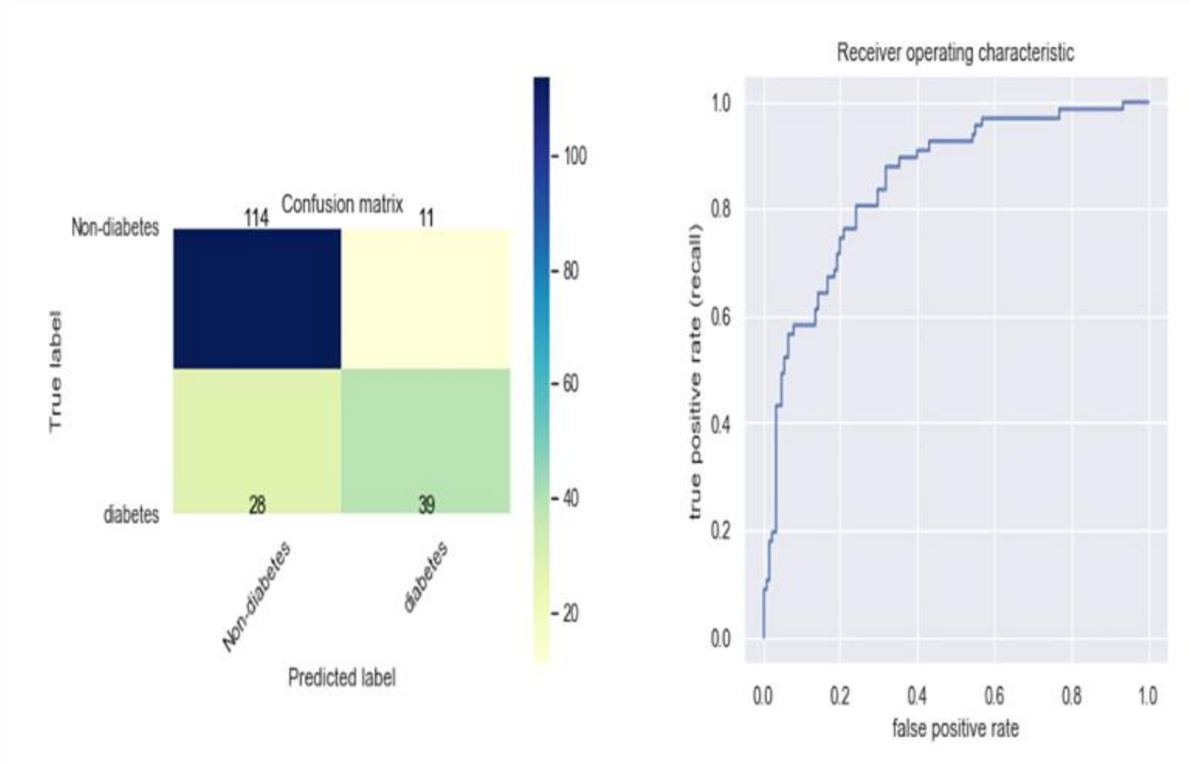

Figure 6 Confusion Matrix and ROC of Random Forest

Extreme Gradient Boosting (XGBoosting)

Extreme Gradient Boosting(XGB) is the second model trained and tested. and shown in Figure 7. From the confusion matrix we can get the accuracy and other performance as mentioned below. The AUC (Area Under Curve) visually appears to be around 0.88–0.92, which is excellent.

Accuracy= (TP +TN)/(TP + TN + FP +FN)

=(51+107)/(51+107+18+16)=82.3%

The precision =(TP)/(TP + FP)

=51/(51+18)=73.9%

The recall =(TP)(TP + FN)

=51/(51+16)=76.1%

True Negative Rate (TNR)=$TN/(TN+FN)$ =107/(107+16)=87.0%

False Positive Rate (FPR) =1−TNR=1−76.9%=13.0%

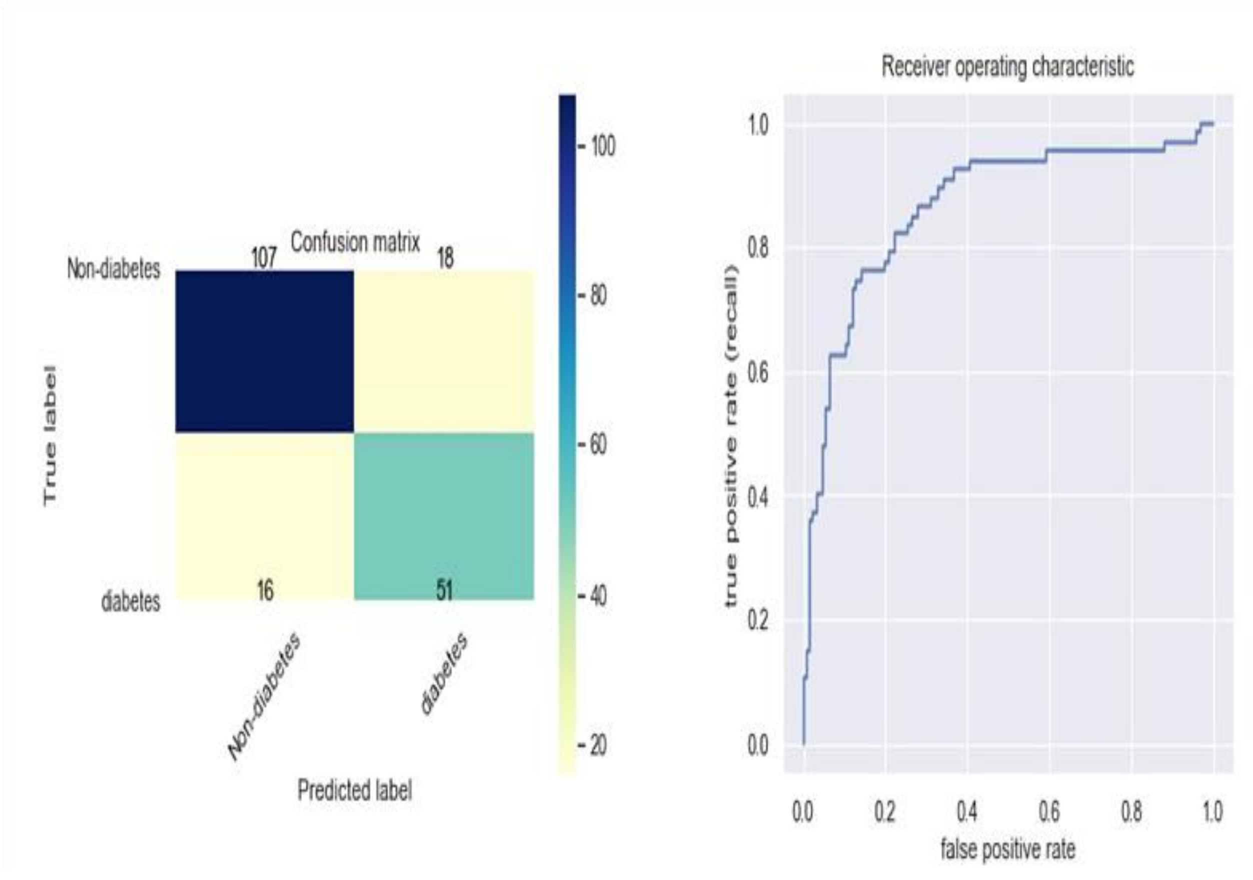

Figure 7 Confusion Matrix and ROC of XGBoost

Model Performance Analysis

The performance scores of RF, XGBoost and Ensemble model are listed below in Table 2. The Ensemble (RF+XGB) model has the best accuracy score (84%), highest recall (0.92), and highest AUC (0.91), as per the table's findings. The recall is sometimes referred as as the percentage of individuals with a condition who have a positive test result is known as sensitivity. Recall is a crucial statistic for model evaluation.

Table 2 Comparison of Model Performance

Model	Accuracy Score	Recall score	Precision	F1 score	Area undercurve (train)	Area under curve(test)
Random Forest	0.7760	0.58	0.78	0.67	0.7441	0.7470
XGBoost	0.8229	0.76	0.74	0.75	0.9293	0.8086
Ensemble RF+XGB	0.84	0.92	0.80	0.86	0.93	0.91

The performance of RF, XGBoost and Ensemble machine learning model with respect to Performance measures is illustrated in following figure 8.

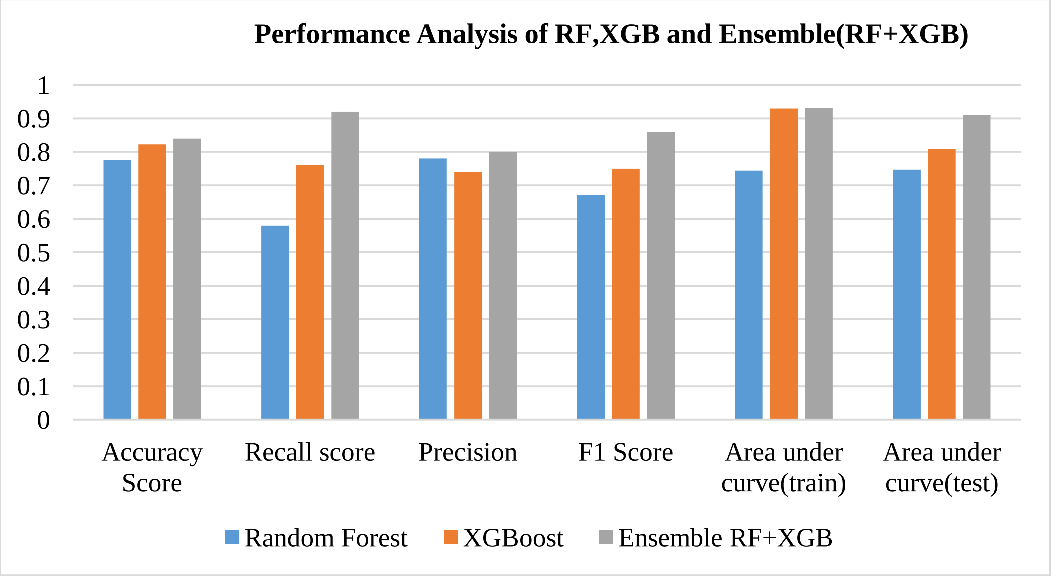

Figure 8 Performance Analysis of Random Forest, XGBoost and Ensemble Algorithm

SHAP plot Analysis

From these SHAP values, it was found that the most influencing features are glucose, age and BMI. The following Figures 9 (a) and (b) present the SHAP summary plots for RF and XGB models.

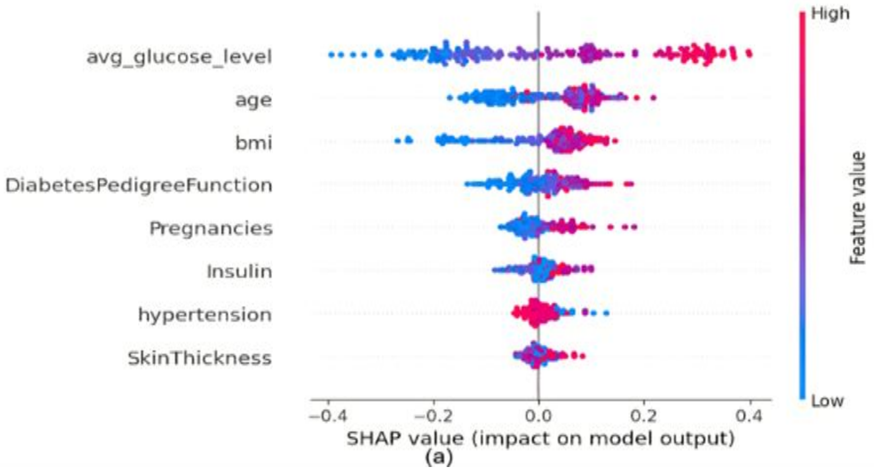

Figure 9 (a) Random Forest with most contributing features as glucose level, age, and BMI

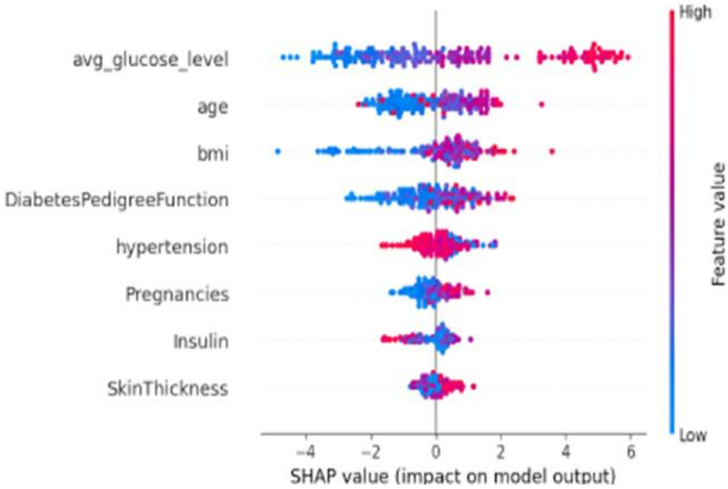

Figure 9 (b) XGBoost with most contributing features glucose level, age, and BMI

Table 3. The most influencing features by SHAP analysis

Model	top 3 features
XGBoost	glucose level, age, BMI
Random Forest	glucose level, age, BMI

Table 3 presents the three most influential features from two applied models and as identified from their respective SHAP summary plots. It is noted that the feature ‘average glucose level’ consistently exerts the highest influence.

Performance Comparison of Ensemble Models

As shown in Table 2, the performance comparison of ensemble models reveals that the stacked Ensemble RF + XGB model performs well across all evaluation metrics, demonstrating strong reliability in classification tasks. In this study, Random Forest (RF) and XGBoost (XGB) act as the base models. The Ensemble RF + XGB model achieves the best results.

Conclusion

This study demonstrates that combining ensemble learning with interpretability techniques like SHAP can significantly enhance diabetes prediction performance. This study investigated the use of Random Forest and XG Boost machine learning classifiers for diabetes prediction. The integration of Random Forest (RF) and XGBoost (XGB) models through a soft voting ensemble achieved outstanding results, with an AUC of 0.91, an accuracy of 0.84, a precision of 0.80, and a recall of 0.92 indicating both strong predictive ability and reliability. The SHAP value analysis revealed that glucose, age, and BMI were the most influential factors contributing to diabetes risk, aligning closely with established clinical understanding. These findings highlight the value of feature interpretability in medical AI, as it allows models to provide not only accurate predictions but also meaningful insights into key health determinants. Comparative analysis further showed that the XGBoost model outperformed the Random Forest model, achieving higher scores across all major performance metrics, including F1-score, accuracy, and recall. Overall, the ensemble approach leveraging XGBoost as a core component, supported by SHAP-based feature interpretation, presents a promising and transparent framework for improving early diabetes detection and supporting clinical decision-making.

Conflict of interest

Author declares no conflict of interest.

References

Ahmed, A., Khan, J., Arsalan, M., Ahmed, K., Shahat, A. A., Alhalmi, A., & Naaz, S. (2025). Machine learning algorithm-based prediction of diabetes among female population using PIMA dataset. *Healthcare, 13*(1), 37.

Bateja, R., Dubey, S. K., & Bhatt, A. K. (2024). Diabetes prediction and recommendation model using machine learning techniques and MapReduce. *Indian Journal of Science and Technology, 17*(26), 2747–2753. https://doi.org/10.17485/IJST/v17i26.530

Chandra Sen, P., Hajra, M., & Ghosh, M. (2020). Supervised classification algorithms in machine learning: A survey and review. *Advances in Intelligent Systems and Computing, 937,* 43–59. https://doi.org/10.1007/978-981-13-7403-6_5

Chou, C.-Y., Hsu, D.-Y., & Chou, C.-H. (2023). Predicting the onset of diabetes with machine learning methods. *Journal of Personalized Medicine, 13*(3), 406. https://doi.org/10.3390/jpm13030406

Dharmarathne, G., Jayasinghe, T. N., Bogahawaththa, M., Meddage, D., & Rathnayake, U. (2024). A novel machine learning approach for diagnosing diabetes with a self-explainable interface. *Healthcare Analytics, 5*, 100301. https://doi.org/10.1016/j.health.2024.100301

Dzira, N. J., Mazdadi, M. I., Farmadi, A., Saragih, T. H., Kartini, D., & Abdullayev, V. (2025). Enhancing diabetes prediction accuracy using random forest and XGBoost with PSO and GA-based feature selection. *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, *7*(2), 295–306.

El-Sofany, H., El-Seoud, S. A., Karam, O. H., El-Latif, Y. M. A., & Taj-Eddin, I. A. T. F. (2024). A proposed technique using machine learning for the prediction of diabetes disease through a mobile app. *International Journal of Intelligent Systems, 2024,* 1–13. https://doi.org/10.1155/2024/6688934

Evwiekpaefe, A. E., Abdulkadir, N., Nigerian Defence Academy, & Nigerian Defence Academy. (2023). A predictive model for diabetes mellitus using machine learning techniques (A study in Nigeria). *The African Journal of Information Systems, 15*(1), 1–1. https://digitalcommons.kennesaw.edu/ajis/vol15/iss1/1

Febrian, M. E., Ferdinan, F. X., Sendani, G. P., Suryanigrum, K. M., & Yunanda, R. (2023). Diabetes prediction using supervised machine learning. *Procedia Computer Science, 216,* 21–30. https://doi.org/10.1016/j.procs.2022.12.107

Gonzalez, U., & Flores, V. (2021). A novel hybrid fusion model of random forest and XGBoost for diabetes prediction. *Journal of Biomedical Engineering, 7*(2), 123–135.

Gowthami, S., Venkata Siva Reddy, R., & Ahmed, M. R. (2024). Exploring the effectiveness of machine learning algorithms for early detection of type 2 diabetes mellitus. *Measurement: Sensors, 31,* 100983. https://doi.org/10.1016/j.measen.2023.100983

Gundogdu, S. (2023). Efficient prediction of early-stage diabetes using XGBoost classifier with random forest feature selection technique. *Multimedia Tools and Applications, 82,* 34163–34181. https://doi.org/10.1007/s11042-023-15165-8

Hasan, M. K., Alam, M. A., Das, D., et al. (2020). Diabetes prediction using ensembling of different machine learning classifiers. *IEEE Access, 8,* 76516–76531.

Ibrahim, A., & Adnan, A. (2021). The role of machine learning algorithms for diagnosing diseases. *Journal of Applied Science and Technology Trends, 2*(1), 10–19. https://doi.org/10.38094/jastt20165

Ismail, L., & Materwala, H. (2025, January 30). IDMPF: Intelligent diabetes mellitus prediction framework using machine learning. *Applied Computing and Informatics, 21*(1–2), 78–89. https://doi.org/10.1108/ACI-10-2020-0094

Kumari, S., Kumar, D., & Mittal, M. (2021). An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. *International Journal of Cognitive Computing in Engineering, 2,* 40–46.

Laxmikant, K., Bhuvaneswari, R., & Natarajan, B. (2023). An efficient approach to detect diabetes using XGBoost classifier. *2023 Winter Summit on Smart Computing and Networks (WiSSCoN)*, 1–8. IEEE.

Modak, S. K. S., & Jha, V. K. (2024). Diabetes prediction model using machine learning techniques. *Multimedia Tools and Applications, 83*(38), 38523–38549. https://doi.org/10.1007/s11042-023-16745-4

Mohanty, P.K.; Francis,S.A.J.; Barik, R.K.; Roy, D.S.; Saikia, M.J. Leveraging Shapley Additive
Explanations for Feature Selection in Ensemble Models for Diabetes Prediction. Bioengineering 2024, 11, 1215. https://doi.org/10.3390/ bioengineering11121215

Mujumdar, A., & Vaidehi, V. (2019). Diabetes prediction using machine learning algorithms. *Procedia Computer Science*, *165*, 292–299. https://doi.org/10.1016/j.procs.2020.01.047

Naz, H., & Ahuja, S. (2020). Deep learning approach for diabetes prediction using PIMA Indian dataset. *Journal of Diabetes & Metabolic Disorders*, *19*(1), 391–403. https://doi.org/10.1007/s40200-020-00520-5

Sarwar, M. A., et al. (2018). Prediction of diabetes using machine learning algorithms in healthcare. *Proceedings of the 24th International Conference on Automation & Computing.* Newcastle University, Newcastle upon

Tyne, United Kingdom.

Shambharkar, S. S., Moon, P. S., & Bainalwar, P. A. (2023). Machine learning-based approach for early detection and prediction of chronic diseases. In *Proceedings of the 1st DMIHER International Conference on Artificial Intelligence in Education and Industry 4.0 (IDICAIEI)* (pp. xx–xx). IEEE. https://doi.org/10.1109/IDICAIEI58380.2023.10406914

Shapiro, M. R., Tallon, E. M., Brown, M. E., & others. (2025). Leveraging artificial intelligence and machine learning to accelerate discovery of disease-modifying therapies in type 1 diabetes. *Diabetologia*, *68*(3), 477–494. https://doi.org/10.1007/s00125-024-06339-6

Shetty, D., Rit, K., Shaikh, S., & Patil, N. (2017). Diabetes disease prediction using data mining. *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*. IEEE. https://doi.org/10.1109/ICIIECS.2017.8275915

Talukder, M. A., Islam, M. M., Uddin, M. A., Kazi, M., Khalid, M., Akhter, A., & Moni, M. A. (2024). Toward reliable diabetes prediction: Innovations in data engineering and machine learning applications. *Digital Health*. Advance online publication. https://doi.org/10.1177/20552076241271867

VijiyaKumar, K., Lavanya, B., Nirmala, I., & Caroline, S. S. (2019). Random forest algorithm for the prediction of diabetes. *2019 International Conference on Systems Computation, Automation and Networking (ICSCAN).*

Xiong, K., Cao, G., Jin, M., & Ye, B. (2025). A multi-modal deep learning approach for predicting type 2 diabetes complications: Early warning system design and implementation. *Journal of Theory and Practice of Engineering Science*, *5*(1), 54–63. https://doi.org/10.53469/jtpes.2025.5(01).06