Intelligent Machine Learning Based Diabetes Prediction System Using Explainable Machine Learning

***Aprajita Singh**
****Aman Prakash Janoriya**

ABSTRACT

This paper presents the implementation and experimental evaluation of an intelligent machine learning-based diabetes prediction system using the Pima Indians Diabetes Dataset. The system is designed around an XGBoost classifier and includes data preprocessing, feature preparation, stratified train-test partitioning, model training, prediction, performance evaluation and feature-importance analysis. The dataset contains records and eight clinical predictors: Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction and Age. Five machine learning models—Logistic Regression, Decision Tree, Random Forest, Support Vector Machine and XGBoost—were implemented for comparative evaluation. The models were evaluated using Accuracy, Precision, Recall, F1-Score, ROC-AUC and Confusion Matrix. XGBoost feature-importance analysis identifies the clinical variables contributing most strongly to prediction.

Keywords:- Diabetes Prediction, XGBoost, Machine Learning, Explainable AI, SHAP, Feature Importance, Classification**.**

*Aprajita Singh, Research Scholar, Department Of Computer Science & Engineering, Radha Raman Engineering College, Bhopal (M.P.), aprajita.71988@gmail.com.
**Aman Prakash Janoriya, Asst. Professor, Department Of Computer Science & Engineering, Radha Raman Engineering College, Bhopal (M.P.), amanprakashjanoriya@gmail.com.

I. INTRODUCTION

Diabetes mellitus is a major chronic disease that requires timely identification and risk assessment. Conventional clinical decision-making depends on multiple measurements and expert interpretation, while machine learning can learn nonlinear relationships between clinical variables and diabetes outcomes from historical data. The supplied research synopsis proposes an intelligent diabetes prediction framework that compares multiple machine learning algorithms and uses explainable artificial intelligence to improve transparency.

XGBoost is selected as the principal implementation model because it is an ensemble gradient-boosting algorithm capable of learning nonlinear relationships through a sequence of decision trees while controlling model complexity through regularization. The implementation therefore follows a pipeline consisting of dataset preparation, preprocessing, feature preparation, train-test partitioning, XGBoost classification, evaluation and feature-importance analysis.

The objective of this implementation is not only to obtain a diabetes prediction but also to compare the predictive performance of different machine learning models and provide interpretable information about the variables influencing the prediction.

II. RELATED WORK

Previous research on diabetes prediction has evaluated statistical classifiers, decision trees, ensemble methods and explainable machine learning. Studies based on the Pima Indians Diabetes Dataset established a benchmark for comparing classification algorithms, while more recent work has emphasized ensemble learning and explainability. The supplied research documents particularly emphasize XGBoost, Random Forest and explainable methods such as SHAP and LIME. The following selected studies position the present implementation

Table1. Selected literature relevant to diabetes prediction and explainable machine learning

Author / Year	**Method**	**Dataset / Domain**	**Major Contribution**
Sisodia & Sisodia (2018)	Classification algorithms	Pima Indians Diabetes	Compared conventional ML classifiers for diabetes prediction.
Friedman (2001)	Gradient Boosting	General ML	Established the gradient boosting framework underlying modern boosting methods.
Breiman (2001)	Random Forest	General ML	Introduced an ensemble of randomized decision trees for robust classification.
Chen & Guestrin (2016)	XGBoost	General ML	Presented scalable and regularized tree boosting with strong

			predictive performance.
Ribeiro et al. (2016)	LIME	General ML	Introduced local model-agnostic explanations for individual predictions.
Lundberg & Lee (2017)	SHAP	General ML	Proposed a unified framework for feature-attribution explanations.
Athanasiou et al. (2020)	Explainable XGBoost	Diabetes-related risk	Demonstrated XGBoost with SHAP for explainable clinical risk assessment.
Naz & Ahuja (2022)	SMOTE + SMO	Pima Diabetes	Addressed diabetes prediction and class imbalance using resampling and classification.
Chang et al. (2023)	ML + SHAP	Pima Diabetes	Combined diabetes prediction with SHAP-based feature interpretation.
Allani (2025)	XAI-based diabetes prediction	Diabetes risk	Presented interactive explainable diabetes-risk prediction using ensemble models and SHAP/LIME.

The literature indicates a transition from conventional classification toward ensemble learning and explainable prediction. These findings motivate the present implementation, which compares baseline and ensemble classifiers while placing XGBoost at the center of the proposed prediction workflow

III. PROPOSED METHODOLOGY

The proposed implementation follows eight major phases: Dataset Collection, Data Preprocessing, Feature Preparation, Train-Test Partitioning, XGBoost Model Construction, Model Evaluation, Feature Importance Analysis and Prediction Result Interpretation.

A. Dataset Description

The experimental dataset contains 768 patient records. Each record includes eight input variables and one binary target variable, Outcome. Outcome = 0 represents a non-diabetic observation and Outcome = 1 represents a diabetic observation.

Table 2. Dataset attributes used in the diabetes prediction system.

Attribute	**Description**
Pregnancies	Number of pregnancies
Glucose	Plasma glucose concentration
BloodPressure	Diastolic blood pressure
SkinThickness	Triceps skin-fold thickness
Insulin	Two-hour serum insulin
BMI	Body Mass Index
DiabetesPedigreeFunction	Diabetes hereditary-risk function
Age	Patient age
Outcome	Target class: 0 = non-diabetic, 1 = diabetic

B. Data Preprocessing

The preprocessing stage checks the dataset for duplicate records, missing observations and invalid measurements. For Glucose, BloodPressure, SkinThickness, Insulin and BMI, a value of zero is treated as a missing-like observation because zero is not physiologically meaningful for these measurements. Median imputation is applied using training-data statistics. Standardization is applied to Logistic Regression and SVM, while tree-based models operate on the imputed feature values.

C. Feature Preparation

The eight clinical predictors are retained as the principal feature set. Feature preparation focuses on data quality rather than generating a large number of derived variables. The resulting feature matrix is supplied to the machine learning classifiers. XGBoost subsequently provides model-based feature importance.

D. Data Partitioning

The dataset is divided into 80% training and 20% testing subsets using stratified sampling with random state 42. The training data are used for model fitting and 10-fold stratified cross-validation, while the held-out test data are used only for final performance reporting.

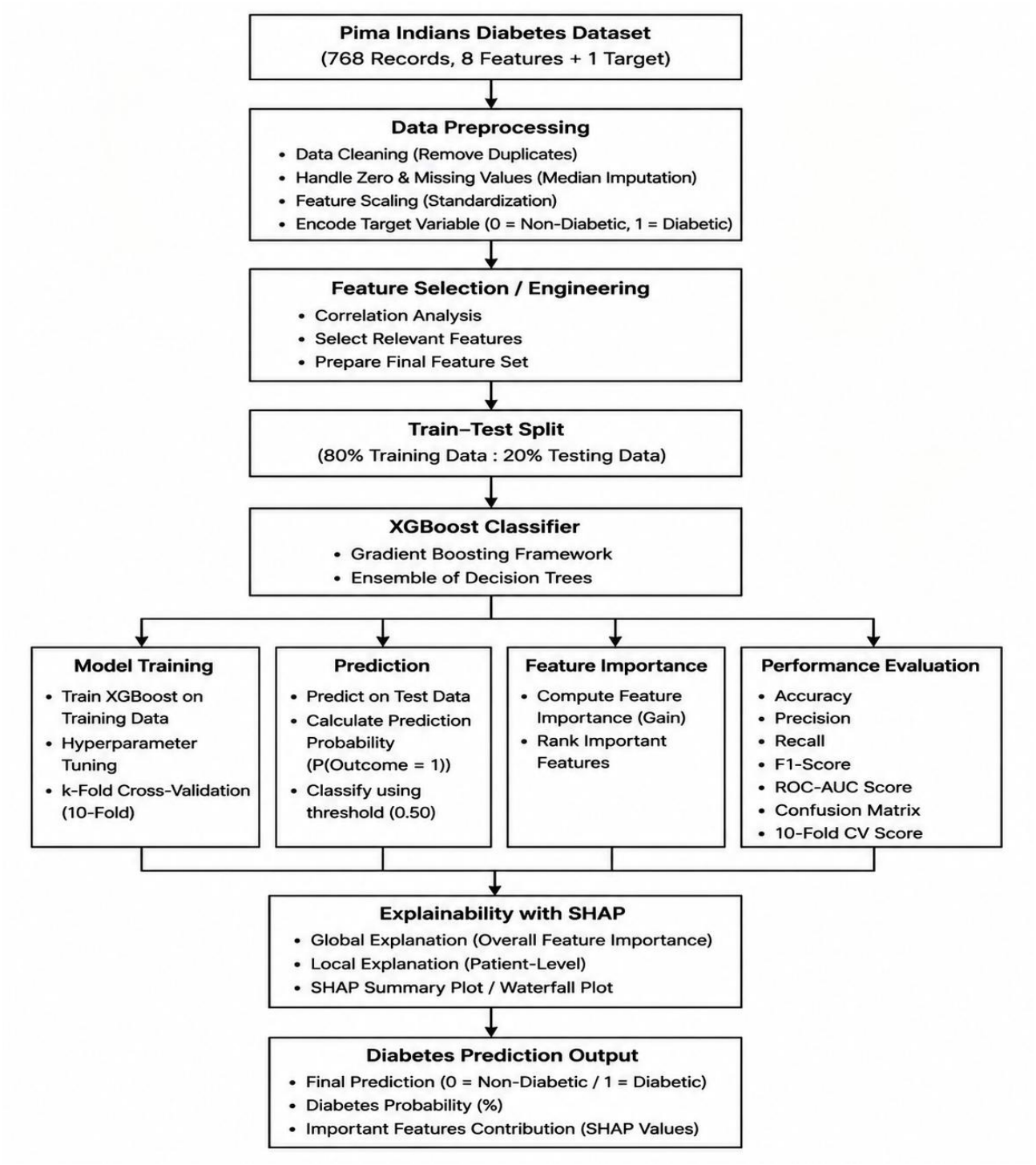

Figure 1: Proposed XGBoost-based intelligent diabetes prediction architecture.

E. XGBoost Model Construction

XGBoost is an optimized gradient boosting framework based on decision trees. Instead of constructing a single classifier, the algorithm sequentially adds trees, where each new tree

attempts to reduce the prediction error of the existing ensemble. For an observation x_i, the ensemble prediction can be represented as:

$$\hat{y}_i = \Sigma(k=1 \text{ to } K)\ f_k(x_i) \quad (1)$$

where f_k represents the kth decision tree and K is the number of trees. The objective function combines the prediction loss with a regularization term that controls tree complexity:

$$Obj(\theta) = L(\theta) + \Omega(\theta) \quad (2)$$

The regularized objective discourages unnecessarily complex trees and helps reduce overfitting. In the present implementation, XGBoost is configured with 300 trees, maximum depth 3, learning rate 0.05, subsample 0.85 and column subsample 0.85. The binary logistic objective is used for diabetes classification.

F. Proposed XGBoost Algorithm

Input: Preprocessed Pima Indians Diabetes Dataset.

Output: Predicted diabetes class and probability.

Step 1: Load the diabetes dataset.

Step 2: Separate the eight input features from the Outcome target.

Step 3: Replace invalid zero values in selected clinical variables with missing values.

Step 4: Apply median imputation using training-data statistics.

Step 5: Split the data into stratified 80% training and 20% testing subsets.

Step 6: Initialize the XGBoost classifier with the selected hyper parameters.

Step 7: Train the ensemble of decision trees on the training data.

Step 8: Generate class predictions and diabetes probabilities for the test data.

Step 9: Compute Accuracy, Precision, Recall, F1-Score and ROC-AUC.

Step 10: Generate the confusion matrix and ROC/Precision–Recall curves.

Step 11: Perform 10-fold stratified cross-validation on the training data.

Step 12: Calculate XGBoost feature importance and identify influential clinical attributes.

Step 13: Return the final diabetes prediction, probability and explanatory feature information.

IV. RESULTS AND DISCUSSION

The proposed implementation was evaluated using the 768-record dataset. The held-out test set contains 154 observations. Five classifiers were evaluated under the same train-test protocol. The objective was to determine whether XGBoost provides a useful balance between predictive accuracy and clinical screening metrics.

Table 3. Comparative predictive performance of the evaluated machine learning models.

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**ROC-AUC**	**10-Fold CV**
Logistic Regression	70.78%	60.00%	50.00%	54.55%	81.30%	78.99% ± 3.94%
Decision Tree	75.97%	63.93%	72.22%	67.83%	76.22%	71.50% ± 3.10%
Random Forest	74.03%	65.91%	53.70%	59.18%	81.63%	76.22% ± 5.36%
SVM	74.03%	65.22%	55.56%	60.00%	79.64%	77.70% ± 4.21%
XGBoost	77.92%	69.23%	66.67%	67.92%	82.57%	75.57% ± 3.70%

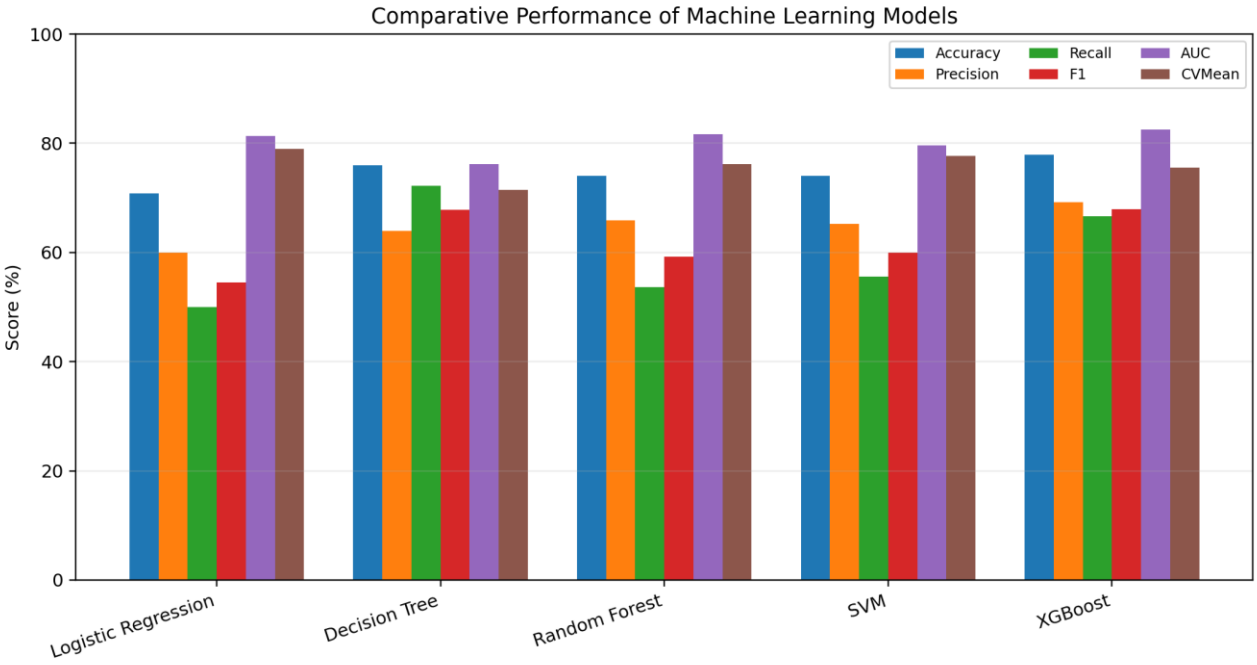

Figure 2: Comparative performance of Logistic Regression, Decision Tree, Random Forest, SVM and XGBoost.

XGBoost achieved the highest held-out test accuracy of 77.92% and the highest ROC-AUC of 82.57%. Decision Tree achieved the highest recall of 72.22%, which is important for identifying diabetic cases. Logistic Regression achieved the highest mean 10-fold cross-validation accuracy of 78.99%. These differences show that model selection should not be based on accuracy alone.

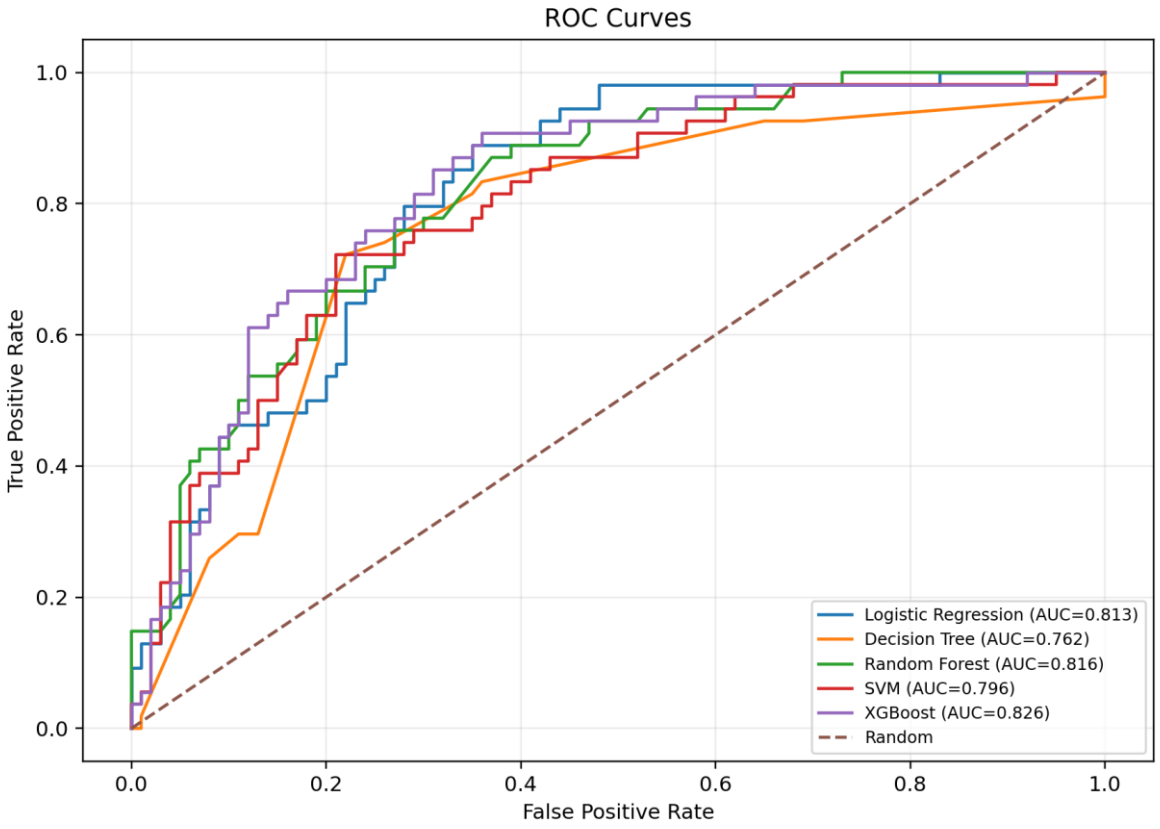

Figure 3: Receiver Operating Characteristic curves for the evaluated models.

ROC analysis shows that XGBoost provides the strongest overall ranking capability, with an AUC of 0.826. Random Forest and Logistic Regression follow with AUC values of 0.816 and 0.813 respectively. The ROC curve is useful for comparing models over a range of classification thresholds rather than at only the default 0.50 threshold.

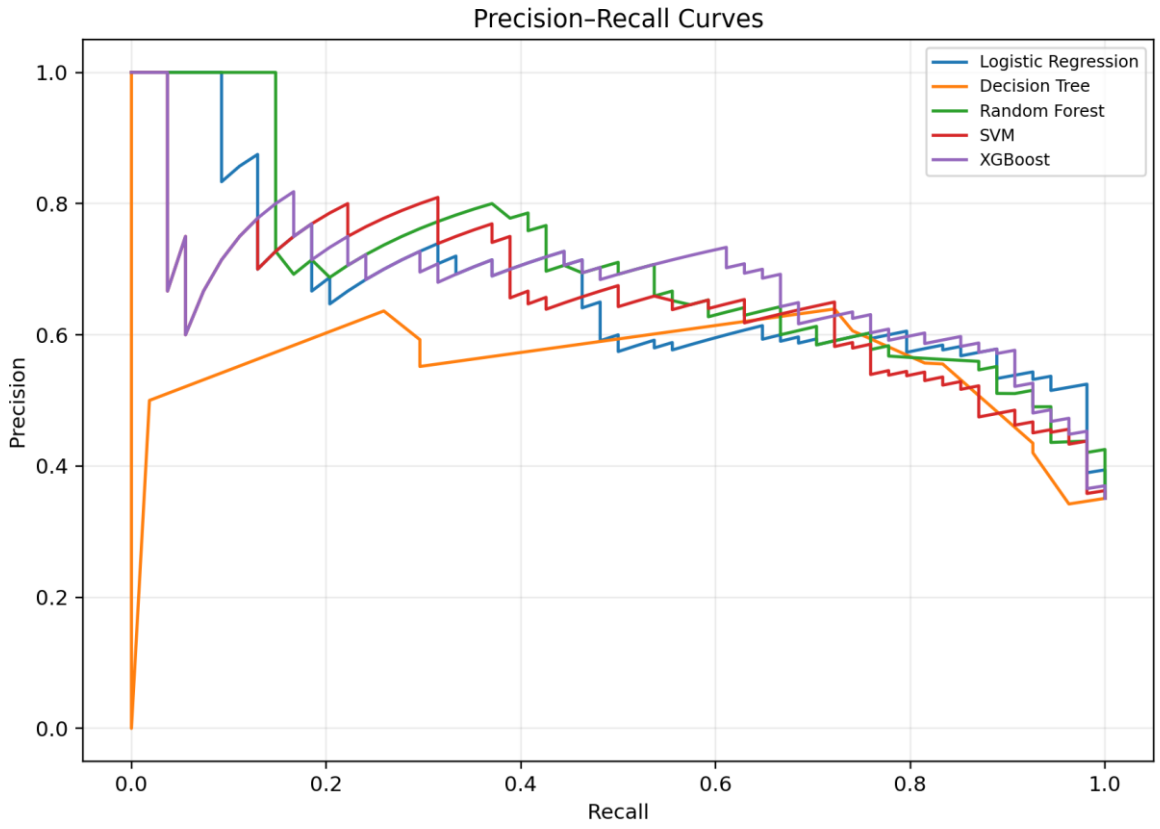

Figure 4: Precision–Recall curves for the evaluated diabetes prediction models.

Precision–Recall analysis provides additional information for the positive diabetic class. The curves demonstrate the trade-off between detecting more diabetic cases and maintaining precision. This analysis is important because the dataset contains fewer diabetic observations than non-diabetic observations.

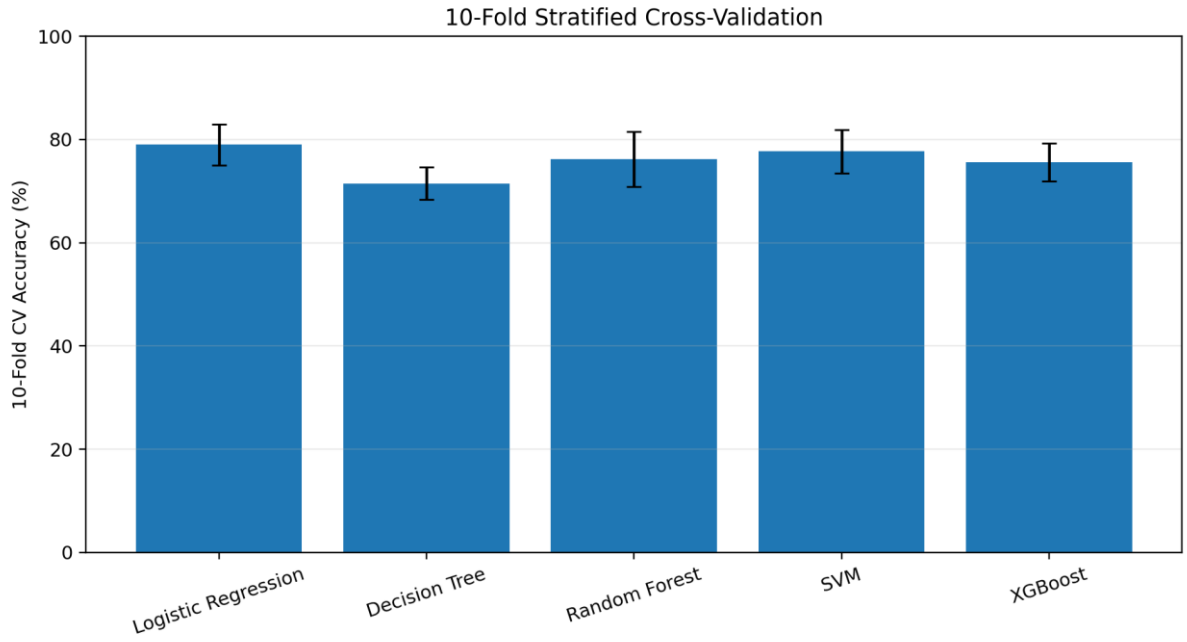

Figure 5: Ten-fold stratified cross-validation accuracy.

Cross-validation demonstrates that Logistic Regression has the highest average validation accuracy, whereas XGBoost has a lower average CV accuracy but stronger held-out test accuracy and ROC-AUC. The result indicates some variability between validation folds and reinforces the importance of reporting both cross-validation and an independent test result.

A. Confusion Matrix Analysis

Table 4. Confusion-matrix components for the evaluated models.

Model	**TN**	**FP**	**FN**	**TP**
Logistic Regression	82	18	27	27
Decision Tree	78	22	15	39
Random Forest	85	15	25	29
SVM	84	16	24	30
XGBoost	84	16	18	36

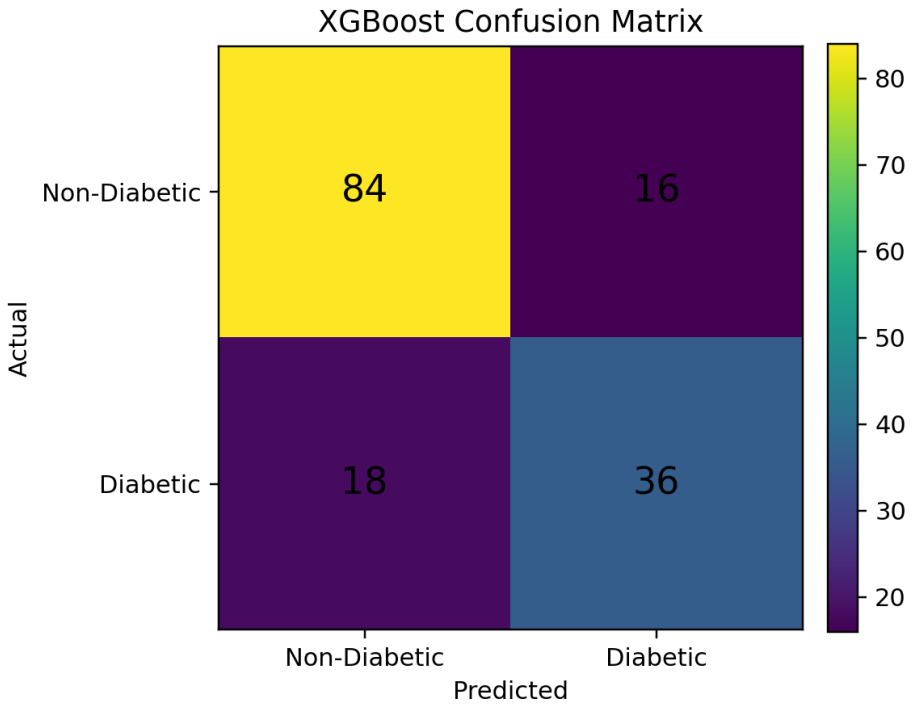

Figure 6: Confusion Matrix of the XGBoost diabetes prediction model.

For XGBoost, the confusion matrix contains 84 true negatives, 16 false positives, 18 false negatives and 36 true positives. The model therefore correctly classifies 120 of the 154 test observations. The 18 false-negative cases are important because they correspond to diabetic observations predicted as non-diabetic, demonstrating why recall should be considered together with accuracy.

B. Feature Importance Analysis

Table 5. XGBoost feature-importance ranking.

Feature	**XGBoost Importance**
Glucose	0.2724
BMI	0.1435
Age	0.1323
Insulin	0.1092

DiabetesPedigreeFunction	0.0969
Pregnancies	0.0901
SkinThickness	0.0804
BloodPressure	0.0752

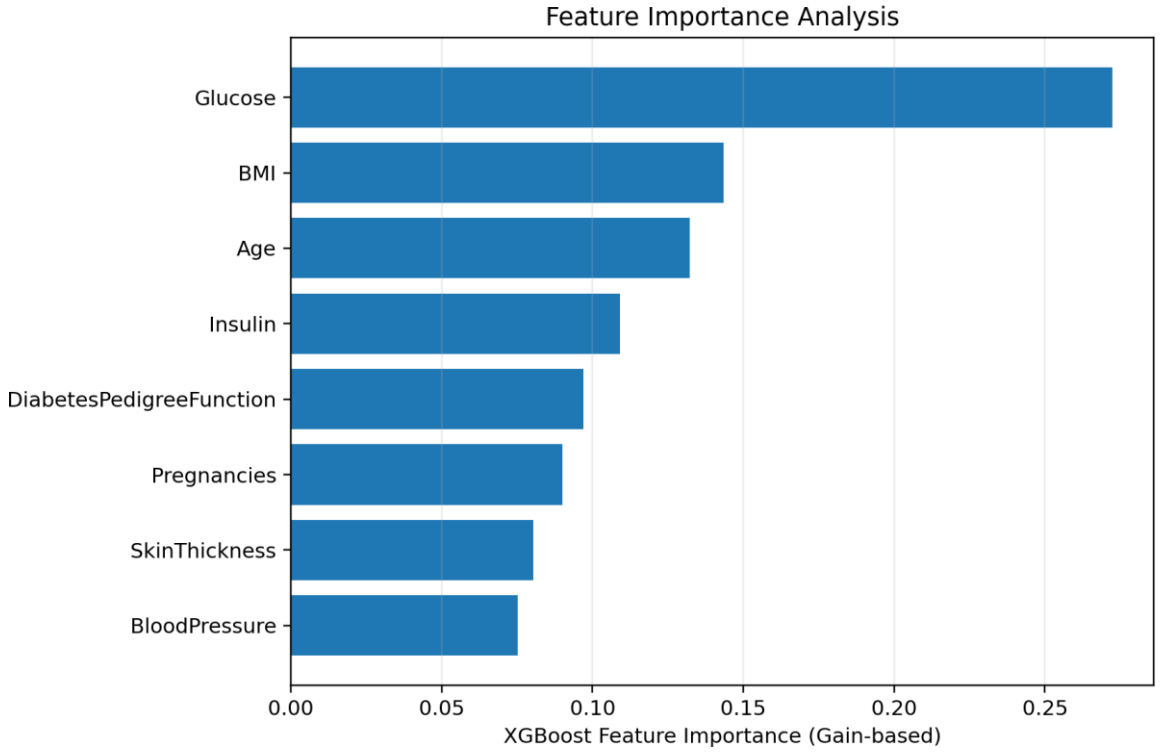

Figure 7: XGBoost feature-importance analysis.

The feature-importance analysis provides a model-specific explanation of which variables contribute most strongly to the XGBoost prediction. Glucose is the dominant clinical predictor in the present implementation, followed by BMI, Age and DiabetesPedigreeFunction. These variables provide a clinically interpretable basis for examining why the model assigns a higher or lower diabetes probability.

C. Comparative Analysis

Table 6. Comparative characteristics of the evaluated algorithms.

Algorithm	**Strengths**	**Limitations**
Logistic Regression	Fast, interpretable baseline	Limited nonlinear modeling
Decision Tree	Simple and interpretable; high recall in this test	Can overfit and is unstable
Random Forest	Robust ensemble and strong ROC-AUC	Larger model and lower interpretability
SVM	Effective nonlinear classifier	Probability estimation and scaling required
XGBoost	High test accuracy, ROC-AUC, regularization and feature importance	Sensitive to hyperparameters and data quality

V. CONCLUSION

This study presented an implementation of an intelligent machine learning-based diabetes prediction system centered on XGBoost and supported by comparative model evaluation and feature-importance analysis. The system used the Pima Indians Diabetes Dataset and applied preprocessing, stratified 80:20 partitioning and 10-fold cross-validation. Among the evaluated models, XGBoost achieved the strongest held-out test accuracy and ROC-AUC, while Decision Tree achieved the highest recall. The results demonstrate that XGBoost is a strong candidate for the proposed diabetes prediction application, although model selection should consider the clinical importance of false negatives and not accuracy alone.

The proposed architecture can be providing patient-level prediction and graphical result presentation. Future work should include larger and more diverse datasets, external validation, probability calibration, threshold optimization, hyper parameter optimization and explicit SHAP/LIME patient-level explanations before considering any clinical deployment.

REFERENCES

1. J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.
2. L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
3. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794, 2016.
4. M. T. Ribeiro, S. Singh and C. Guestrin, "Why Should I Trust You?: Explaining the Predictions of Any Classifier," Proc. KDD, 2016.
5. S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," Advances in Neural Information Processing Systems, 2017.
6. Sisodia and Sisodia, "Prediction of Diabetes Using Classification Algorithms," 2018.
7. G. Athanasiou et al., "An Explainable XGBoost-Based Approach Towards Assessing the Risk of Cardiovascular Disease in Patients with Type 2 Diabetes Mellitus," 2020.
8. Naz and Ahuja, "SMOTE-SMO-Based Expert System for Type II Diabetes Detection Using PIMA Dataset," 2022.
9. Chang, Chen and Chen, "e-Diagnosis System for Diabetes Prediction Using Machine Learning and SHAP," 2023.
10. Chang et al., "Pima Indians Diabetes Mellitus Classification Based on Machine Learning Algorithms," 2023.
11. Allani, "Interactive Diabetes Risk Prediction Using Explainable Machine Learning," 2025.
12. Alkhanbouli et al., "The Role of Explainable Artificial Intelligence in Disease Prediction: A Systematic Review," 2025.