Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Comparative Evaluation of Machine Learning Models for Diabetes Prediction with Threshold Optimization and Explainable AI

Munsi Rakibul Islam
Department of Computer Science and Engineering
Netaji Subhash Engineering College
Kolkata, India

Sugata Chanda
Department of Computer Science and Engineering
Netaji Subhash Engineering College
Kolkata, India

Soumi Ghosh
Department of Computer Science and Engineering
Netaji Subhash Engineering College
Kolkata, India

Dr. Chandra Das
Department of Computer Science and Engineering
Netaji Subhash Engineering College
Kolkata, India

Dr. Shilpi Bose
Department of Computer Science and Engineering
Netaji Subhash Engineering College
Kolkata, India

Abstract*—Timely prediction of diabetes is crucial for prompt management and better outcomes. In this paper, we perform comparative analysis of six machine learning algorithms including Logistic Regression, Random Forest, Voting Classifier, Stacking Classifier, AdaBoost, and CatBoost on a large-scale diabetes database with nearly 100,000 patients' information. In order to solve the problem of class imbalance and improve the performance of the models, class-weighted learning, threshold adjustment, and five-fold cross validation were introduced to the proposed evaluation method. The models' performance was estimated using such metrics as Precision, Recall, F1-score, and ROC-AUC. The best discriminative performance was shown by CatBoost with the ROC-AUC of 0.978. Threshold adjustment helped to increase the F1-score of CatBoost from 0.685 to 0.807 and its precision from 0.563 to 0.951 without compromising the recall too much. Five-fold cross validation proved the reliability of the models with the mean ROC-AUC of 0.9777 and the standard deviation of 0.0007. SHapley Additive exPlanations (SHAP) analysis revealed the following four features as the most important ones: HbA1c level, blood glucose level, age, and BMI.

***Keywords*—Diabetes prediction, machine learning, CatBoost, ensemble learning, threshold optimization, explainable artificial intelligence, SHAP, ROC-AUC..**

I. Introduction

Diabetes is a growing concern among the number of non-communicable diseases that both high- and low- and middle-income countries struggle with. The main reason is that it is asymptomatic at early stages and manifests at late stages when significant damage to organs occurs. It is therefore important to detect it at early stages through the screening process. At the same time, screening requires laboratory tests that are not always available to populations. As electronic health records and population-level datasets have grown in size and granularity, researchers have turned to machine learning as a complementary lens, one capable of surfacing risk signals embedded in combinations of variables that are rarely flagged by single-biomarker analysis [8]. In spite of the significant body of published research, the generalizability of the results to different patient populations is far from settled: performance heavily depends on the exact choice of the dataset and methodology, and it is often challenging to identify a particular approach that consistently exhibits its predictive power across various settings [8]. At the same time, robustness and interpretability are frequently overlooked, with a model that performs consistently across the test set being less valuable than one that can generalize to other, real-world populations. This paper attempted to address these issues by evaluating six classifiers on a 100,000-sample pool of demographical and lifestyle-related data that includes the patients' biometric stats, assessing how the choice of threshold, train-test split, and the use of SHAP values impact the model performance. It is hypothesized that each metric can be utilized to identify the most relevant features, determine the optimal threshold for each classifier, compare the performance between individual models, and highlight the most significant improvements brought upon by explainable AI methods.

II. Related Work

Machine learning techniques have been widely applied for diabetes prediction due to their ability to identify patterns from clinical and demographic data. A recurring theme in the diabetes prediction literature is that algorithm choice alone rarely determines outcome quality. Studies that applied systematic preprocessing pipelines — addressing class imbalance, normalizing continuous features, and tuning hyperparameters via grid or Bayesian search — generally reported larger gains than those that switched between algorithms on otherwise unprocessed data [1], [2], [3]. This pattern suggests that the upstream data preparation stage deserves at least as much attention as the model architecture itself.

Boosting-based classifiers have attracted particular interest within this domain. Gradient Boosting, XGBoost, LightGBM, AdaBoost, and CatBoost have each been evaluated for diabetes classification, and comparative results consistently place them above simpler linear models on datasets where clinical variables interact in non-obvious ways [1]. Glucose and HbA1c measurements, for instance, do not independently explain diabetes risk — their predictive value depends heavily on how they interact with BMI, age, and comorbidity status, relationships that tree-ensemble structures are well suited to capture without manual feature engineering. Studies have reported that boosting-based models frequently

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

achieve higher predictive accuracy than individual classifiers when supported by appropriate preprocessing and parameter optimization [1], [4].

Beyond single boosting algorithms, stacking architectures have also been tested for this task. Rather than relying on one dominant learner, these approaches route the outputs of several base models into a meta-classifier that learns to reconcile their disagreements [2]. Ahmed et al. found that this two-tier design yielded more stable recall on diabetic minority-class samples than any individual base model achieved on its own, an advantage most apparent under cross-validation where single-model results tended to fluctuate more across folds [2]. Comparative studies indicate that ensemble strategies often provide more stable performance than standalone machine learning algorithms, particularly when evaluated using cross-validation techniques [2], [5].

Systematic reviews of this field identify two persistent gaps that technical improvements in accuracy have not resolved [3], [5]. First, the PIMA Indians Diabetes Database — a 768-record cohort drawn from a specific demographic — continues to be used as the primary benchmark in a disproportionate share of published studies [11], raising legitimate questions about whether findings generalize to ethnically and clinically diverse populations [3], [5]. Second, most studies treat the default 0.5 classification threshold as fixed and stop short of integrating explainability tools that would allow clinicians to audit individual predictions; these two issues together limit the real-world utility of otherwise well-performing models.

Motivated by these observations, the present study performs a comparative evaluation of multiple machine learning models using a large-scale diabetes prediction dataset. The study further investigates threshold optimization, train-test split comparison, cross-validation-based performance assessment, and SHAP-based explainability [7] to provide both predictive effectiveness and model interpretability.

III. Dataset and Preprocessing

A. Dataset Description

The diabetes prediction data used for this research was collected from a public Kaggle dataset [10]. The data includes the attributes of diabetes risk. Initially, the data had 100,000 instances and nine features, including the target attribute.

TABLE I. VARIABLE DESCRIPTION OF THE DIABETES PREDICTION DATASET

Name	Type, Subtype	Description
Gender	Categorical, Nominal	Gender of the individual
Age	Numeric, Continuous	Age of the individual in years
Hypertension	Categorical, Binary	Hypertension status (0 = No. 1 = Yes)
Heart_Disease	Categorical, Binary	Heart disease status (0 = No, 1 = Yes)
Smoking_History	Categorical. Nominal	Smoking behaviour category
BMI	Numeric, Continuous	Body Mass Index
HbA1c_Level	Numeric, Continuous	Glycated hemoglobin measurement
Blood_Glucose_Level	Numeric, Continuous	Blood glucose concentration
Diabetes	Categorical, Binary	Target variable (0 = Non-diabetic, 1 = Diabetic)

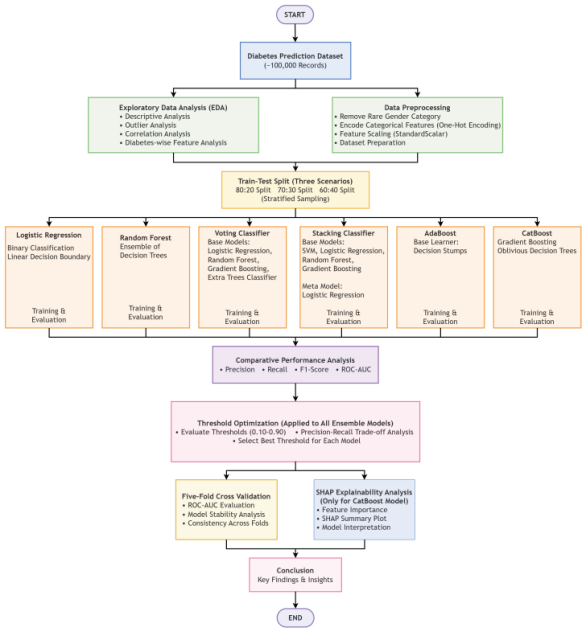

Fig. 1. Proposed methodology framework.

The dataset contains eight predictor variables and one target variable. A summary of the features used in this study is presented in Table I. The diabetes attribute serves as the target variable for binary classification.

Inspection of the dataset revealed that there were no missing observations across all nine attributes, which simplified the data preparation process as no data imputations were necessary. However, an inspection of the response distribution showed there was a significant imbalance between the two classes: while the dataset contained 100,000 instances, only 8.5% of them were classified as “diabetic”. This meant that any classifier would have a substantial bias towards predicting negative cases (“not diabetic”), which would lead to unrealistically high levels of accuracy. In order to address this issue, the class weights for the optimizer were adjusted in accordance with Eq. (1) so that the costs associated with misclassification errors for the positive class (“diabetic”) were increased in order to balance the optimization criteria.

$$w_i = \frac{N}{K \times n_i} \tag{1}$$

where:

w_i = weight assigned to class i

N = total number of samples

K = number of classes

n_i = number of samples belonging to class i

Class weights were computed inversely proportional to class frequencies, where minority-class observations received higher weights than majority-class observations. This approach helps improve sensitivity toward diabetic cases while preserving overall predictive performance.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

B. Data Cleaning and Feature Preparation

An exploratory analysis was performed to examine feature distributions, category frequencies, and potential data quality issues. The gender attribute contained a rare category labeled "Other," representing only 18 records. To reduce noise and avoid instability caused by extremely low-frequency categories, these records were removed from the dataset. The smoking_history attribute contained categories such as "No Info" and "not current". Although these values may introduce some ambiguity, they represented a substantial portion of the dataset and were therefore retained during preprocessing. This decision allowed the models to learn potential patterns associated with these categories rather than removing a large number of observations.

After cleaning, the dataset contained 99,982 records. Categorical variables were encoded into numerical representations to facilitate machine learning model training. Continuous variables, including age, BMI, HbA1c level, and blood glucose level, were retained as numerical features.

C. Exploratory Data Analysis

Distributional analysis of the features painted a clear picture: blood glucose and HbA1c concentrations were markedly elevated in records labelled diabetic, with relatively little overlap in their upper tails compared to non-diabetic records. Age and BMI followed a similar pattern — diabetic individuals skewed older and heavier — though with considerably more distributional overlap, suggesting these two variables contribute context rather than serving as standalone discriminators. Correlation analysis confirmed this hierarchy, with glucose and HbA1c posting the strongest positive associations with the target label. These empirical observations were later reaffirmed by SHAP attribution values derived from the final model [7].

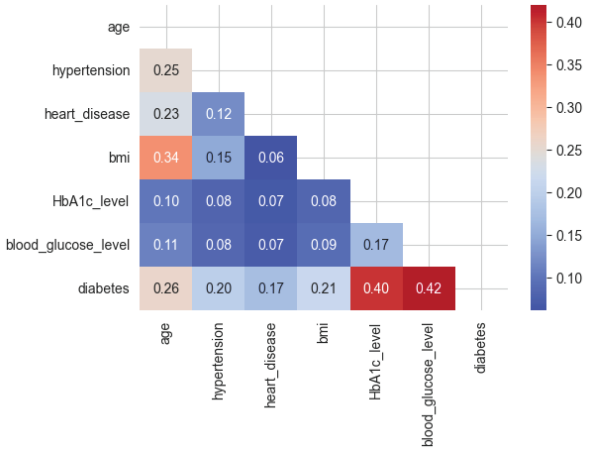

Fig. 2. Correlation Heatmap

Correlation analysis further revealed that blood glucose level and HbA1c level possessed the highest positive correlations with the target variable, indicating their importance for diabetes prediction. These observations were further supported through SHAP-based feature importance analysis [7] of the CatBoost model, which achieved the highest overall predictive performance.

IV. Methodology

A. Experimental Setup

The cleaned data was split into training and test sets at the ratio of 80:20. The training set was used for model training, whereas, the test set was used for model evaluation. The 70:30 and 60:40 splits were also considered in order to assess the effect on the model performance.

Logistic regression, random forest, voting classifier, stacking classifier, AdaBoost, and CatBoost [6] are considered. The voting classifier uses logistic regression, random forest, and gradient boosting to combine various logistic regression results. It uses soft voting to determine the final class, which implies utilizing the predicted class probability for voting to identify the most frequent prediction. The stacking classifier uses support vector machine (SVM), logistic regression, random forest, and gradient boosting as base classifiers, whereas logistic regression is used as a meta-classifier to allow combining different classifiers for achieving better performance. Such models were selected since they represent traditional machine learning algorithms, ensemble learning, and boosting techniques, which are applied in health care prediction problems..

B. Model Evaluation Metrics

Model performance was evaluated using Precision, Recall, F1-Score, and Receiver Operating Characteristic Area Under Curve (ROC-AUC).

Four metrics were used in combination because no single measure adequately characterizes performance under class imbalance. Precision captures how reliably a positive prediction is correct — a critical consideration given that false alarms carry their own clinical and resource costs. Recall reflects how thoroughly the model retrieves true diabetic cases, which is equally important since missed diagnoses delay intervention. The F1-score effectively combines the two metrics by down weighting the extremes. The ROC-AUC supplements this type of analysis by capturing the separation ability at all levels of probability, which is relevant if one has to rely on different operating points of the model. Given the dataset's 91.5% non-diabetic majority, overall accuracy was deliberately excluded as a primary metric, as it would reward trivial majority-class prediction. Because the dataset exhibits class imbalance, emphasis was placed on F1-Score and ROC-AUC rather than overall accuracy.

C. Threshold Optimization

Binary classifiers output a continuous probability score, and the conventional practice of converting that score into a class label at 0.5 reflects a computational convenience rather than a clinical optimum. In an imbalanced medical setting, a lower threshold tends to increase recall at the expense of precision, while a higher threshold does the reverse—and the right trade-off depends on the relative consequences of false negatives versus false positives in that specific context. To identify thresholds that better serve this problem, the probability outputs of the Voting Classifier, Stacking Classifier, AdaBoost, and CatBoost models were evaluated across a range of decision thresholds, with precision, recall, and F1-score tracked at each point. The Threshold–F1 curves were then analyzed to identify stable high-performance regions, and candidate thresholds were compared based on their precision–recall trade-off. The final threshold for each model was selected by achieving a balanced combination of

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

precision, recall, and F1-score rather than solely maximizing a single evaluation metric.

D. Cross-Validation

To move beyond point estimates from a single train-test partition, five-fold cross-validation was applied to all models. Each fold cycle reserved one fifth of the data exclusively for validation while training on the remaining four fifths, rotating until every record had served once in the held-out role. Reporting both the mean ROC-AUC and its standard deviation across the five iterations provided a cleaner picture of each model's stability: a high mean paired with low variance indicates a model whose performance does not depend heavily on which records happen to land in the test set.

E. Explainable AI using SHAP

Post-hoc interpretability was evaluated using the SHAP (SHapley Additive exPlanations) [7], a game-theoretic approach, which explains the output of a given model by quantifying the contribution of each feature to the prediction result. Unlike the global interpretation methods, which quantify the importance of each variable on the model's performance, the SHAP algorithm calculates the marginal contribution of each variable to a specific prediction. Such an approach better suits medical data analysis because it allows understanding the interplay between the input variables and the outcome on a case-by-case basis. In other words, the SHAP analysis helps identify the variables, which positively or negatively contributed to the prediction outcome for a particular patient. The post-hoc interpretability was conducted for the best-performing CatBoost model to elucidate the comparative study's result..

V. Experimental Results and Discussion

A. Comparative Performance Analysis

The predictive performance of Logistic Regression, Random Forest, Voting Classifier, Stacking Classifier, AdaBoost and CatBoost was evaluated using Precision, Recall, F1-Score, and ROC-AUC metrics. The results obtained using the 80:20 train-test split are presented in Table II.

TABLE II. Performance Comparison of Machine Learning Models

Model	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.42	0.89	0.57	0.962
Random Forest	0.94	0.69	0.80	0.963
Voting Classifier	0.90	0.71	0.79	0.972
Stacking Classifier	0.92	0.71	0.80	0.975
AdaBoost	0.99	0.68	0.80	0.975
CatBoost	0.56	0.87	0.68	0.978

CatBoost algorithm has the highest ROC-AUC score among others, reaching 0,978, which proves its accuracy in terms of distinguishing between the positive and negative classes. Thus, the default model of CatBoost has high recall (87,3%), while its precision is slightly lower than the precision of other models.

The Voting, Random Forest, Stacking Classifier and AdaBoost algorithms have a similar performance, with a ROC-AUC score higher than 0,96 and an F1-score around 0,8. The Logistic Regression has a moderate recall but a low precision and F1-score compared to ensemble algorithms. The results indicate that ensemble learning and boosting approaches are more effective than traditional linear classification methods for diabetes prediction on the selected dataset.

To further compare the discriminative ability of the best-performing models, ROC curves for all evaluated models are shown in Fig. 3.

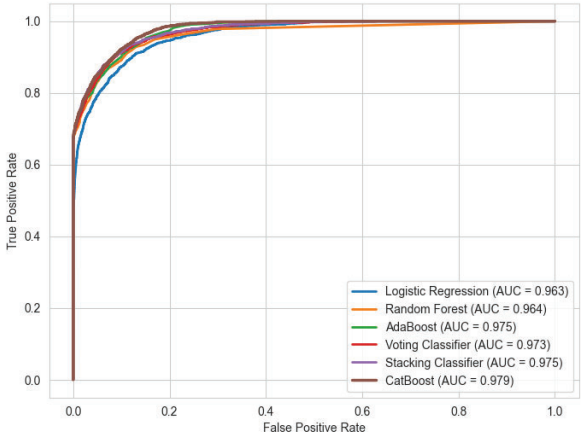

Fig. 3. ROC Curve Comparison of all six models

The ROC curves indicate strong classification capability for all evaluated models. However, the CatBoost curve consistently remains above the other models across most threshold values, demonstrating superior discriminative performance for diabetes prediction.

B. Impact of Train-Test Split Selection

To evaluate the stability of model performance, experiments were conducted using three train-test split configurations: 80:20, 70:30, and 60:40.

TABLE III. Performance under Different Train-Test Splits

Model	Split	Precision	Recall	F1	ROC-AUC
Logistic Regression	80:20	0.42	0.89	0.57	0.963
	70:30	0.43	0.88	0.57	0.962
	60:40	0.42	0.88	0.57	0.963
Random Forest	80:20	0.94	0.69	0.80	0.964
	70:30	0.94	0.69	0.80	0.962
	60:40	0.94	0.69	0.80	0.964
Voting Classifier	80:20	0.90	0.71	0.79	0.973
	70:30	0.91	0.71	0.80	0.972
	60:40	0.90	0.71	0.80	0.973
Stacking Classifier	80:20	0.92	0.71	0.80	0.975
	70:30	0.92	0.71	0.80	0.975
	60:40	0.92	0.72	0.81	0.975
AdaBoost	80:20	0.99	0.68	0.80	0.975
	70:30	0.99	0.68	0.80	0.975
	60:40	0.98	0.68	0.80	0.976
CatBoost	80:20	0.56	0.87	0.68	0.979

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

	70:30	0.58	0.87	0.69	0.979
	60:40	0.57	0.87	0.69	0.979

The obtained results showed insignificant differences between the different split ratios. In particular, all the considered models were characterized by comparable ROC-AUC and F1-scores for different splits. It can be concluded that the considered machine learning models are rather stable and insensitive to the size of the training dataset..

The stability of performance across multiple train-test splits further strengthens confidence in the reliability of the developed prediction framework.

C. Threshold Optimization Analysis

The default classification threshold of 0.50 was further analyzed to determine whether improved predictive performance could be achieved through threshold optimization across the evaluated ensemble models.

$$\hat{y} = \begin{cases} 1, p \geq t \\ 0, p < t \end{cases} \quad (2)$$

where:

$\hat{y}$ = predicted class label

p = predicted probability of diabetes

t = decision threshold

A probability threshold (t) was applied to convert predicted probabilities into class labels. While the default threshold is generally set to 0.50, multiple threshold values were evaluated for the Voting Classifier, Stacking Classifier, AdaBoost, and CatBoost models.

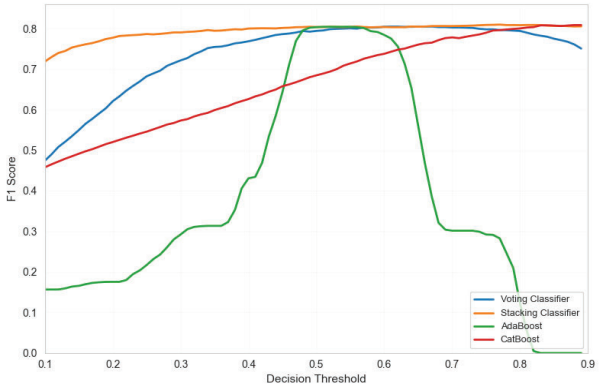

Fig. 4. Threshold–F1 curves for the evaluated ensemble models

Figure 4 shows that the evaluated ensemble models exhibit different threshold-response characteristics, indicating that a common decision threshold is not optimal for all classifiers. Based on these observations, model-specific thresholds were selected for subsequent performance evaluation.

TABLE IV. THRESHOLD OPTIMIZATION COMPARISON

Model	Threshold	Precision (B/A[a])	Recall (B/A[a])	F1 (B/A[a])
Voting	0.60	0.90→0.97	0.71→0.69	0.79→0.80
Stacking	0.70	0.92→0.97	0.72→0.69	0.80→0.81
AdaBoost	0.48	0.99→0.86	0.68→0.74	0.80→0.80
CatBoost	0.85	0.56→0.95	0.87→0.70	0.68→0.81

[a] B/A denotes performance before and after threshold optimization

The selected operating thresholds were 0.60 for Voting Classifier, 0.70 for Stacking Classifier, 0.48 for AdaBoost, and 0.85 for CatBoost. Table IV shows that threshold optimization altered the precision–recall trade-off for each ensemble model without substantially affecting overall predictive performance. CatBoost achieved the largest improvement in F1-score, while the remaining ensemble models attained better-balanced operating points for diabetes prediction.

D. Cross-Validation Analysis

To assess generalization capability and model stability, five-fold cross-validation was performed using ROC-AUC as the evaluation metric. The average ROC-AUC score and standard deviation for each model are presented in Table V.

TABLE V. FIVE-FOLD CROSS-VALIDATION RESULTS

Model	Mean ROC-AUC	Std. Dev
Logistic Regression	0.9618	0.0012
Random Forest	0.9613	0.0017
Voting Classifier	0.9716	0.0017
Stacking Classifier	0.9738	0.0017
AdaBoost	0.9741	0.0007
CatBoost	0.9777	0.0007

The highest mean value of the metric (ROC-AUC score) was achieved by the CatBoost algorithm: 0.9777, with a standard deviation (SD) equal to 0.0007, which indicates the algorithm's high performance. The AdaBoost, Stacking Classifier, and Voting Classifier also demonstrated a relatively high mean value of the metric (ROC-AUC score) over 0.97. The SD for all the algorithms was close to 0, which indicates their stable performance and insignificant dependence on the choice of training and validation samples..

VI. EXPLAINABILITY ANALYSIS

A single metric, prediction score, is of limited value in the clinical context; practitioners need to know which patient characteristics contributed to a certain risk score for the result to be useful in practice, and regulators are now demanding that diagnostic models using AI be transparent and interpretable rather than operating as "black boxes" [9]. The best-performing CatBoost model was therefore subjected to SHAP (SHapley Additive exPlanations) analysis in order to assess the relative contribution of individual predictors.

The two variables that showed the largest difference between the cases were HbA1c and blood glucose, which were the predictors with the highest mean absolute SHAP values, and thus had the greatest impact on the model's output. In other words, the concentration of these substances was the main lever that the model used to differentiate between the two classes, diabetic and non-diabetic.. Age and BMI occupied a secondary tier — individually meaningful contributors, but with narrower attribution ranges than the glycemic markers. Smoking history and hypertension status generated moderate SHAP values, consistent with their roles as contextual risk modifiers rather than primary signals. Gender and heart disease contributed the least across the patient cohort, though individual records with atypical profiles occasionally registered non-negligible attributions for these variables. Notably, this feature hierarchy aligned closely with the distributional differences observed during exploratory analysis, which provides some assurance that the model

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

learned clinically coherent patterns rather than spurious correlations in the training data.

The feature importance identified through SHAP was consistent with the patterns observed during exploratory data analysis. Both HbA1c level and blood glucose level exhibited strong associations with diabetes occurrence and maintained their importance during model training. This consistency increases confidence that the model is learning clinically meaningful relationships rather than relying on spurious patterns.

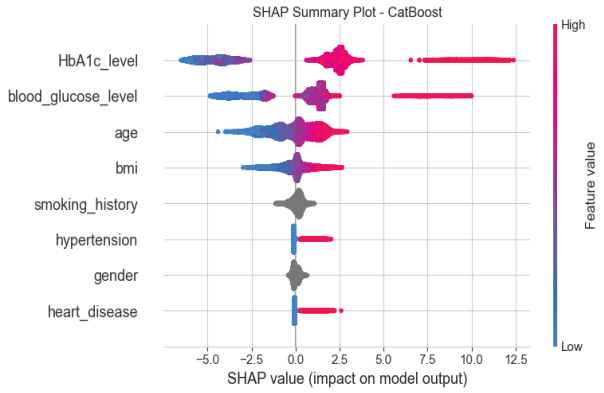

Fig. 5. SHAP summary plot showing feature contributions

Figure 5 presents the SHAP summary plot, illustrating the contribution of individual features to model predictions. Higher HbA1c and blood glucose values generally increased the likelihood of diabetes prediction, whereas lower values contributed toward non-diabetic classifications.

Taken together, the SHAP results confirm that the model's discriminative power is grounded in the same clinical variables that domain knowledge would identify as meaningful — a reassuring sign that strong ROC-AUC numbers reflect genuine learned structure rather than dataset artefacts, and one that strengthens the case for deploying such a framework in settings where prediction must be paired with explanation.

VII. Conclusion

The current work carried out a comparison of numerous machine learning models for the task of diabetes prediction on the largest diabetes prediction dataset consisting of 100k entries. The following classifiers were evaluated both in terms of Precision, Recall, F1-Score, and ROC-AUC metrics: Logistic Regression, Random Forest, Voting Classifier, Stacking Classifier, AdaBoost and CatBoost.

The results of the experiments conducted revealed that CatBoost yielded the best performance with the ROC-AUC metric reaching 0.978 . Overall, the experiments demonstrated that the choice of the optimal threshold for each classifier improved the precision-recall performance of all tested ensemble models; with CatBoost showing the highest improvement in terms of the F1-Score. These findings highlight the importance of threshold selection for imbalanced healthcare datasets.

Model robustness was further confirmed through five-fold cross-validation, where CatBoost demonstrated the highest mean ROC-AUC score with consistently low variation across validation folds. Explainability analysis using SHAP identified HbA1c level, blood glucose level, age, and BMI as the most influential factors contributing to diabetes prediction, aligning with observations obtained during exploratory data analysis.

The results from this study suggest that the value of any individual component—a well-tuned ensemble, calibrated decision thresholds, or SHAP-based feature attribution—is amplified when they are used together. Although CatBoost achieved the strongest discriminative performance, threshold optimization further improved the operating characteristics of the evaluated ensemble models by providing more suitable precision–recall trade-offs for diabetes prediction. Cross-validation and train-test split comparison confirmed that the observed performance was not an artefact of a favourable data split. And SHAP attribution showed that the model's internal logic tracked clinical reality rather than noise. This convergence of evidence points toward an evaluation framework that goes beyond single-metric leaderboard comparisons and treats robustness, calibration, and transparency as first-class requirements alongside accuracy.

References

[1] M. A. R. Chowdhury et al., “A gradient boosting-based framework for diabetes prediction using PIMA Indians dataset,” Frontiers in Genetics, vol. 14, Art. no. 1252159, 2023, doi: 10.3389/fgene.2023.1252159.

[2] H. Ahmed et al., “Improving diabetes disease patients classification using stacking ensemble method with PIMA and local healthcare data,” Heliyon, vol. 10, no. 3, e26142, 2024, doi: 10.1016/j.heliyon.2024.e26142.

[3] A. A. I. Aljawarneh et al., “Machine learning-based diabetes diagnosis using improved quality data preprocessing and feature selection techniques,” Computers in Biology and Medicine, vol. 175, Art. no. 107489, 2025, doi: 10.1016/j.compbiomed.2025.107489.

[4] H. Naz and S. Ahuja, “Deep learning approach for diabetes prediction using PIMA Indian dataset,” Journal of Diabetes & Metabolic Disorders, vol. 19, no. 1, pp. 391–403, 2020, doi: 10.1007/s40200-020-00520-5.

[5] P. B. Khokhar, C. Gravino, and F. Palomba, “Advances in artificial intelligence for diabetes prediction: insights from a systematic literature review,” Artificial Intelligence in Medicine, vol. 164, Art. no. 103132, 2025, doi: 10.1016/j.artmed.2025.103132.

[6] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," Advances in Neural Information Processing Systems (NeurIPS), vol. 31, 2018.

[7] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017.

[8] Q. An, S. Rahman, J. Zhou, and J. J. Kang, "A Comprehensive Review on Machine Learning in Healthcare Industry: Classification, Restrictions, Opportunities and Challenges," Sensors, vol. 23, no. 9, Art. no. 4178, 2023, doi: 10.3390/s23094178.

[9] Z. Sadeghi et al., "A review of Explainable Artificial Intelligence in healthcare," Computers and Electrical Engineering, vol. 118, Art. no. 109370, 2024, doi: 10.1016/j.compeleceng.2024.109370.

[10] Diabetes Prediction Dataset, Kaggle Dataset. Available: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

[11] PIMA Indians Diabetes Dataset, UCI Machine Learning Repository. Available: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database