Science in Information Technology Letters
Vol. 6., No. 2, November 2025, pp. 124-146
ISSN 2722-4139
http://pubs2.ascee.org/index.php/sitech

An interpretable machine learning framework for diabetes classification on imbalanced clinical data

Cicin Hardiyanti P [a,1,*], Avrillaila Akbar Harahap [b,2]

[a] Informatics Department, Universitas Alma Ata, Daerah Istimewa Yogyakarta, Indonesia
[b] Information Systems, Universitas Alma Ata, Daerah Istimewa Yogyakarta, Indonesia
[1] cicinhardiyanti@almaata.ac.id; [2] avrillaila@almaata.ac.id
* Corresponding Author

ARTICLE INFO

Article history
Received September 20, 2025
Revised October 16, 2025
Accepted November 2, 2025

Keywords
Interpretable machine learning
Diabetes prediction
Class imbalance
XGBoost
SHAP

ABSTRACT

Diabetes mellitus is a chronic disease with increasing global prevalence and is frequently diagnosed at advanced stages due to minimal early symptoms. Early detection based on clinical data is essential to prevent long-term complications. However, machine learning-based diabetes classification remains challenged by class imbalance, complex clinical features, and limited interpretability. This study proposes an integrated, leakage-aware classification pipeline that combines Synthetic Minority Oversampling Technique (SMOTE), Boruta feature selection, and XGBoost modelling within a 5-fold stratified cross-validation framework. The dataset consists of 100,000 patient records with a 8.5% prevalence of diabetes. Evaluation focused on metrics suitable for imbalanced data, particularly Recall and Precision-Recall AUC (PR-AUC). On the hold-out test set, the proposed model achieved a recall of 0.75, ROC-AUC of 0.977, and PR-AUC of 0.878. Cross-validation results showed stable performance (recall 0.740 ± 0.019; PR-AUC 0.872 ± 0.009). SHAP-based interpretation identified HbA1c, blood glucose, age, and BMI as dominant predictors, with contribution patterns consistent with established clinical evidence. These results demonstrate that the proposed framework provides accurate and clinically interpretable diabetes risk prediction suitable for screening-oriented applications. Nevertheless, the study is limited by the use of a single-source, cross-sectional dataset without external validation, and further multi-centre and longitudinal evaluation is required to confirm generalizability.

1. Introduction

Diabetes mellitus is a chronic, non-communicable disease whose incidence continues to increase annually. The increasing number of diabetes cases is a serious health problem, especially in developing countries with limited healthcare resources. Diabetes often develops silently without visible signs in the early stages, leading patients to delay diagnosis until complications develop. Therefore, early detection of diabetes is a critical strategy to prevent further complications and help maintain patients' quality of life.

Recent advances in healthcare information systems have enabled the application of Machine Learning (ML) techniques for disease prediction, including diabetes classification. ML algorithms can process large-scale clinical data and model complex nonlinear relationships that traditional statistical approaches may miss. Various algorithms have been used in diabetes classification research, including Logistic Regression [1], Decision Trees [2]–[4], Random Forests [5]–[7], Multilayer Perceptrons (MLP) [8], [9], and other methods.

Despite these advances, several methodological challenges remain. First, clinical diabetes datasets are typically highly imbalanced, where non-diabetic cases substantially outnumber diabetic cases [10], [11]. Under such conditions, models may achieve high accuracy while failing to adequately detect the minority class, which is clinically the most critical group [12], [13]. Therefore, relying solely on accuracy may produce misleading conclusions in medical classification tasks.

Second, clinical datasets often contain features that are redundant, weakly relevant, or highly correlated [14]. Including all available variables without systematic selection may increase model complexity, reduce generalisation performance, and complicate interpretation [15]. Several studies have performed model parameter optimisation or explored certain architectures, such as XGBoost optimisation and variations in neural network structures [16], [17], however, there is still a lack of something that systematically applies feature selection based on statistical relevance to ensure that the features used truly contribute to diabetes prediction.

Third, interpretability has become an essential requirement in healthcare applications. High predictive performance alone is insufficient if clinicians cannot understand the reasoning behind model predictions. Explainable Artificial Intelligence (XAI) techniques, such as Shapley Additive exPlanations (SHAP), have been increasingly used to interpret complex models [12], [13]. However, in many studies, interpretability is treated as an auxiliary step rather than as an integral component of a systematically designed modelling pipeline.

Previous research has shown significant progress in diabetes classification using Machine Learning, particularly with the XGBoost algorithm and ensemble approaches [12], [17], [18], but still has limitations. Studies that use XGBoost algorithms and explainability techniques have generally not explicitly addressed issues of data imbalance or advanced feature selection. On the other hand, research that achieves very high accuracy sometimes lacks adequate validation or uses the same training and test data, which can potentially lead to overfitting and reduce the validity of the results. Some studies are also limited to homogeneous datasets, and generalising models remains a challenge [12].

Based on this issue, a research gap can be identified: the lack of studies developing a diabetes classification pipeline that is fully integrated by combining data imbalance handling, robust feature selection, high-performance classification models, and model interpretability. Specifically, the use of Synthetic Minority Oversampling Techniques (SMOTE) to address class imbalance, combined with Boruta-based feature selection and an XGBoost model, has not yet been widely studied together within a single systematic framework. Furthermore, model evaluation on imbalanced data should still focus on metrics that are more clinically relevant, such as recall, ROC-AUC, and Precision-Recall AUC (PR-AUC), rather than just accuracy.

The main contributions of this research include: 1) The development of a clinically-oriented XGBoost-based framework for imbalanced diabetes classification; 2) The use of appropriate evaluation metrics for imbalanced data; 3) Model interpretation based on SHAP to identify clinical factors that significantly contribute to diabetes prediction.

2. Method

2.1. Research Stages

This research was conducted through an integrated pipeline that starts with clinical data preprocessing, including checking for missing values, encoding categorical features, and splitting the data into training and test sets using a stratified split (80:20) to maintain class proportions. Because the dataset has a significant class imbalance, SMOTE was applied to the training data to improve the model's sensitivity to the diabetes class. Next, feature selection was conducted using the Boruta algorithm to identify the most relevant variables, ensuring only significant features were used in modelling. The classification model was then built using XGBoost with a Repeated Stratified 5-Fold Cross-Validation scheme to ensure stability and generalisation capability. Performance evaluation was conducted using metrics relevant for imbalanced data, namely Recall, ROC-AUC, and PR-AUC, and analysed through the confusion matrix and ROC-PR curves on a separate test set. Finally, the model's interpretability was analysed using SHAP to explain feature contributions both globally and individually, so that the model not only has high predictive performance but is also transparent and explainable. The complete research workflow is shown in Fig. 1.

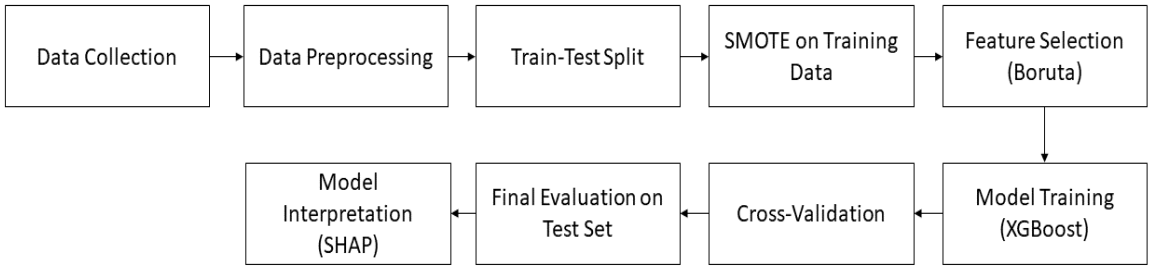


Fig. 1. Research Stages

2.2. Preprocessing Data

The data preprocessing stage is carried out to ensure the dataset is high-quality and ready for modelling. This process aims to minimise potential bias and improve the stability of the classification model. In this study, preprocessing includes checking for missing values, transforming categorical features into numerical form, and splitting the data into training and testing sets using a stratified split to maintain class proportions [19], [20].

2.2.1. Handling Missing Values

Handling missing values is an initial step in clinical data preprocessing because their presence can affect the performance and stability of classification models [21], [22]. In this study, the number of missing values for each feature was examined to ensure data completeness, as shown in Table 1.

The analysis results showed that there were no missing values in all the features used. However, an imputation procedure was still prepared as a precautionary measure using mean-based imputation for

numerical features. The dataset was then confirmed to be free of missing values and ready for the next stage.

Table 1. Number of Missing Values in Each Feature

Feature	Number of Missing
gender	0
age	0
hypertension	0
heart_disease	0
smoking_history	0
bmi	0
HbA1c_level	0
blood_glucose_level	0
diabetes	0

Table 1 shows that all features have good data completeness, so no samples need to be removed due to missing values.

2.2.2. Encoding Categorical Features

The dataset used contains several categorical features that machine learning algorithms cannot directly process. Therefore, a process was performed to convert categorical features into numeric values using label encoding, as shown in Table 2. The categorical features in the dataset are gender and smoking history. The label encoding process maps each category to a unique numeric value. This approach was chosen because it is simple and suitable for a relatively limited number of categories.

Table 2. Categorise Feature and Encoding Method

Feature	Primitive Data Type	Number of Categories	Encoding Method
gender	Categorical	3	Label Encoding
smoking_history	Categorical	6	Label Encoding

After the encoding process, all features are numerical, allowing them to be used for feature selection and classification.

2.2.3. Separation of Training Data and Test Data

After preprocessing, the dataset is split into training and test sets. The division is carried out using an 80% training and 20% test split, with a stratified approach to maintain the proportions of the diabetes and non-diabetes classes in both data subsets, as shown in Table 3. This approach aims to ensure that the test data reflect the actual class distribution, making model performance evaluation more objective and realistic.

Table 3. Training and Testing Data Distribution

Data Subset	Sample Size	Non-Diabetic	Diabetes
Training (80%)	80.000	±73.200	±6.800
Testing (20%)	20.000	±18.300	±1.700
Total	100.000	91.500	8.500

2.3. Handling Data Imbalance (SMOTE)

In the dataset used in this study, the class proportions are significantly imbalanced between the non-diabetes and diabetes classes, as shown in Table 4. This condition can lead the model to achieve high overall accuracy. However, it has a low ability to detect patients with diabetes, who are the most important class in a medical context.

Table 4. Class Distribution before Handling Imbalance

Class	Sample Size	Presentase
Non-Diabetes (0)	±91.500	±91.5%
Diabetes (1)	±8.500	±8.5%
Total	100.000	100%

This distribution shows that the dataset has a significant class imbalance, so standard classification approaches are likely to yield models biased toward the majority class. To address class imbalance, this study uses the Synthetic Minority Oversampling Technique (SMOTE). Unlike simple oversampling methods that duplicate minority data, SMOTE creates new samples by interpolating between neighbouring minority data points.

2.3.1. Application of SMOTE on Training Data

To prevent data leakage, SMOTE was applied exclusively within each training fold during cross-validation. In each fold, the training subset was separated from the validation subset, and SMOTE was fitted and applied only to the training data. The validation subset retained the original class distribution and was never involved in the resampling process. No resampling was performed on the independent test set.

SMOTE generates synthetic minority samples by interpolating between neighbouring minority instances, thereby balancing the class distribution in the training data [23]–[25]. This strategy enables the model to capture better the characteristics of the minority class (diabetes) and improves sensitivity (recall) without inflating performance estimates.

2.4. Boruta Feature Selection

This study applies a feature selection stage to identify variables that are truly relevant to diabetes prediction. Feature selection is performed after preprocessing and before training the classification model, so only the selected features are used in the subsequent modelling stage.

Boruta is a wrapper-based feature selection method that uses the Random Forest algorithm to evaluate the importance of each feature [26]–[29]. The main principle of Boruta is to compare the importance of original features with that of randomly generated shadow features. A feature is considered important if its importance value is consistently higher than that of the shadow features. With this approach, Boruta can identify all features relevant to the target variable, whether linear or nonlinear. To prevent supervised data leakage, Boruta is applied independently within each training fold during cross-validation. Specifically: 1) Boruta is fitted only using the training data in that fold; 2) Relevant features are selected based solely on the training data; 3) The selected feature set is then applied to the validation data without refitting. At no stage is the validation data used to determine feature relevance. This nested

feature selection approach ensures that the reported performance metrics reflect the true generalisation capability.

2.4.1. Implementation of Boruta on the Research Dataset

The Boruta algorithm is applied using a Random Forest Classifier with a weighting scheme to account for class imbalance. The feature selection process is performed on all features generated by preprocessing. The use of Boruta feature selection provides several key advantages, including: 1) Reducing model complexity without losing important information; 2) Improving the stability of model performance on test data; 3) Facilitating the interpretation of results, especially when combined with interpretability methods such as SHAP. This study uses only high-relevance features, and the XGBoost model built in the subsequent stage is expected to deliver better predictive performance and more meaningful clinical interpretation.

2.5. Classification Modelling using XGBoost

Extreme Gradient Boosting (XGBoost) is an ensemble learning algorithm based on gradient-boosted decision trees, widely used for modelling tabular data, including clinical data. XGBoost has strong capabilities for capturing nonlinear relationships among variables, handling complex feature interactions, and providing effective regularisation mechanisms to reduce the risk of overfitting. In diabetes classification, the heterogeneous and imbalanced nature of the data makes XGBoost an appropriate choice, as it is robust to variations in data distribution and delivers stable predictive performance.

2.5.1. Model Architecture and Parameters Used

The XGBoost model in this study used features selected by Boruta, namely age, BMI, HbA1c, and blood_glucose_level. The use of selected features aims to ensure that the model learns only the most relevant information for diabetes prediction.

The main parameters used in model training are shown in Table 5. These parameter settings were chosen to balance model complexity and generalisation.

Table 5. XGBoost Model Parameters

Parameter	Value
n_estimators	150 (CV) / 200 (final model)
max_depth	6
learning_rate	0.05
objective	binary: logistic
eval_metric	logloss
random_state	42

2.5.2. Model Training Scheme

Model training was conducted using a Repeated Stratified K-Fold Cross-Validation approach with five folds (5-fold). This approach ensures that the distribution of diabetes and non-diabetes classes remains proportional across folds, resulting in a more stable evaluation of model performance that is not biased toward any particular subset of data. In each fold, the following steps were carried out sequentially: 1) Splitting the data into training and test sets based on the fold indices; 2) Applying SMOTE only to

the training data to address class imbalance; 3) Training the XGBoost model using the SMOTE-processed training data; 4) Evaluating the model's performance using the test data without resampling.

The exclusive application of SMOTE on the training data aims to prevent data leakage and maintain the validity of model evaluation. The XGBoost model was trained on data optimised through two key stages: handling class imbalance with SMOTE and feature selection with the Boruta algorithm. The combination of these two approaches enables the model to be more sensitive in recognising diabetes cases as the minority class, avoid learning from irrelevant features, and reduce model complexity without compromising predictive performance.

This approach establishes a systematic, integrated modelling pipeline for diabetes classification. Overall, the modelling stages in this study can be summarised as follows: 1) Classification models are built using XGBoost; 2) Training data is optimised through SMOTE to address class imbalance; 3) Only features resulting from Boruta selection are used in model training; 4) Validation is conducted using a stratified cross-validation scheme.

This step serves as the basis for evaluating the model's performance, as discussed in the next section, including the analysis of evaluation metrics and the interpretation of prediction results using the SHAP approach.

2.6. Model Evaluation

Evaluating the performance of a classification model on clinical data with an imbalanced class distribution requires metrics that accurately reflect the model's performance for the minority class. In this study, the evaluation metrics focused on Recall, ROC-AUC, and Precision-Recall AUC (PR-AUC), as these three are more relevant than accuracy alone in the medical context [30].

Recall measures a model's ability to identify patients who actually have diabetes correctly. Recall is defined as the ratio of the number of true positive predictions to the total number of actual positive cases [31]. In the medical field, recall plays an important role because misclassifying diabetic patients as non-diabetic can lead to delayed diagnosis and increase the risk of serious health complications [32].

ROC-AUC is used to measure the model's ability to distinguish between the diabetes and non-diabetes classes overall across various decision thresholds [33]. An ROC-AUC value close to 1 indicates a very good level of class separability.

Precision-Recall AUC (PR-AUC) is a more representative metric for imbalanced data because it specifically emphasises model performance on the positive class (diabetes) [34]. This metric differs from ROC-AUC because it is more sensitive to changes in performance for the minority class.

2.6.1. Model Interpretability Using SHAP

In the field of health, especially in predicting chronic diseases such as diabetes, model accuracy alone is not sufficient to ensure the acceptance and use of Machine Learning-based systems. Medical professionals need a clear understanding of the reasons behind a prediction so that the model's results can be trusted and used as a basis for clinical decision-making. Therefore, interpretability becomes an important aspect in developing diabetes prediction systems that are not only accurate but also transparent and explainable.

This study uses Shapley Additive exPlanations (SHAP) to improve model interpretability. SHAP is a game-theoretic approach that estimates the contribution of each feature to the model's predictions by comparing the change in output when a feature is included or excluded [35]–[37]. SHAP was chosen

because it has several key advantages, namely: 1) Provides mathematically consistent interpretations; 2) Can be applied to complex models such as XGBoost; 3) Capable of explaining predictions both globally and individually.

Global interpretation is also used to identify the features that are overall most influential in predicting diabetes across the entire dataset. In this study, global interpretation was conducted using the SHAP Summary Plot and the SHAP Feature Importance Plot [38]–[40]. These plots show the distribution of SHAP values for each feature, allowing the ranking of feature importance, the magnitude of each feature's contribution to the prediction, and the variance of feature influence across all samples to be observed.

In addition to global interpretation, SHAP is also used to explain predictions at the individual level (Local Explanation) [41]. This approach allows analysis of how specific feature values in a patient contribute to increases or decreases in the predicted probability of diabetes. Local interpretation is highly relevant in a clinical context because it enables healthcare professionals to understand the reasons behind the model's prediction in a specific patient's case, rather than relying solely on general population patterns.

In this study, SHAP was integrated as the final step of the modelling pipeline, after the XGBoost model was trained on data processed by class imbalance handling and feature selection. This approach ensures that interpretation is performed on the final, optimised, and evaluated model.

3. Results and Discussion

3.1. Results of Data Imbalance Handling and Feature Selection

3.1.1. Class Distribution Before and After SMOTE Implementation

Before the modelling process, the class distribution in the training data is analysed to assess the imbalance between the diabetes and non-diabetes classes, as shown in Table 6. Significant class imbalance can bias the model toward the majority class and reduce its ability to detect the minority class. Therefore, a comparison of class distributions is made before and after SMOTE is implemented to evaluate the oversampling technique's effectiveness in balancing the training data.

Table 6. Class Distribution in Training Data Before and After SMOTE

Training Data Condition	Non-Diabetes	Diabetes
Before SMOTE	±73.200	±6.800
After SMOTE	±73.200	±73.200

Applying SMOTE produces a balanced class distribution in the training data, allowing the model to learn patterns from both classes more fairly [42]. With a more balanced class distribution, the training data becomes more representative for the model's learning process. This condition is particularly important in the medical context, as errors in detecting diabetic patients (false negatives) can have serious consequences due to delays in clinical treatment. To clarify the impact of SMOTE, a visualisation of the class distribution before and after oversampling was conducted, as shown in Fig. 2. This visualisation shows a significant difference in class proportions before SMOTE and a more balanced distribution after SMOTE is applied to the training data.

(a) Class distribution before SMOTE

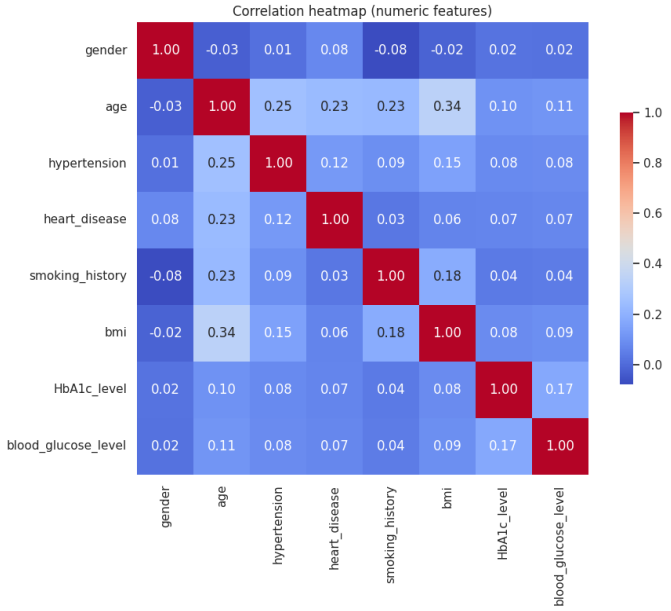
(b) Class distribution after SMOTE

Fig. 2. Class distribution before and after SMOTE (bar chart)

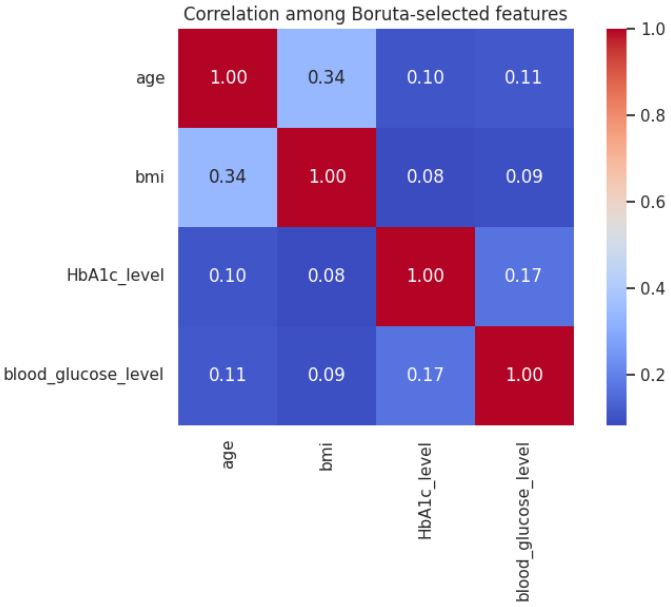
These results indicate that SMOTE successfully addressed class imbalance without removing the original data from the majority class. With a more balanced class distribution, the model becomes more sensitive to patterns representing diabetes patients.

3.1.2. Feature Selection Results Using the Boruta Algorithm

Before feature selection, the dataset contained 8 predictor variables, including demographic and clinical variables. Although all of these features are potentially informative, using all features without selection can increase model complexity and the risk of overfitting. The Boruta algorithm was applied to identify features that are statistically relevant to the diabetes target variable. The selection results showed that, out of the initial eight features, Boruta identified four main features that make significant contributions to diabetes prediction: age, Body Mass Index, HbA1c level, and Blood glucose level as shown in Fig. 3. Meanwhile, the other features were deemed irrelevant and removed from subsequent modelling processes.



(a) Correlation heatmap of all features



(b) Correlation heatmap of selected features

Fig. 3. Correlation heatmap between Boruta features

These results indicate that metabolic clinical indicators play a more dominant role compared to demographic factors or other disease histories. The exclusion of other features demonstrates that Boruta can filter out less relevant or redundant features, retaining only the most important information.

3.2. Performance Evaluation Based on Key Clinical Metrics

To ensure the model's performance is not dependent on a specific data split, the evaluation was conducted using 5-fold Stratified Cross-Validation as shown in Table 7. This approach maintains the proportions of the diabetes and non-diabetes classes in each fold, making the evaluation results more stable and better representative of the characteristics of imbalanced data. For each fold, the XGBoost

model was trained on data that had undergone SMOTE and Boruta feature selection, then evaluated on the test data using Recall, ROC-AUC, and PR-AUC metrics.

Table 7. Model Evaluation Results on Each Fold

Fold	Accuracy	Precision	Recall	F1-Score	ROC-AUC	MCC	PR-AUC
1	0.9640	0.8235	0.7329	0.7756	0.9750	0.7576	0.8691
2	0.9636	0.7974	0.7665	0.7816	0.9774	0.7620	0.8809
3	0.9652	0.8229	0.7518	0.7857	0.9781	0.7677	0.8800
4	0.9597	0.7719	0.7465	0.7590	0.9739	0.7371	0.8663
5	0.9620	0.7972	0.7424	0.7688	0.9766	0.7487	0.8752

Although accuracy is relatively high across all folds, the analysis focuses on Recall and PR-AUC, which are more relevant for clinical data with imbalanced class distributions.

Table 8 summarises the mean and standard deviation of the evaluation metrics obtained from 5-fold stratified cross-validation using the XGBoost + SMOTE model. The consistently high ROC-AUC and PR-AUC values, along with low standard deviations across all metrics, indicate stable, reliable performance in handling imbalanced diabetes classification.

Table 8. Average and Standard Deviation of Evaluation Metrics

Metric	Mean	Standard Deviation
Accuracy	0.9627	±0.0031
Precision	0.8085	±0.0404
Recall	0.7396	±0.0192
F1-Score	0.7716	±0.0147
ROC-AUC	0.9757	±0.0018
MCC	0.7528	±0.0177
PR-AUC	0.8720	±0.0085

The relatively small standard deviation values for most evaluation metrics indicate stable model performance across the five cross-validation folds. In particular, ROC-AUC (±0.0018), PR-AUC (±0.0085), and accuracy (±0.0031) demonstrate consistent discrimination capability and classification stability.

However, precision shows comparatively higher variability (±0.0404). This variation is expected in imbalanced datasets, where small fluctuations in the number of false positives across folds can substantially affect precision. Despite this variability, recall (±0.0192) and F1-score (±0.0147) remain relatively stable, indicating consistent performance in detecting the minority class.

Overall, the low variance observed in discrimination metrics suggests that the SMOTE–Boruta–XGBoost pipeline generalises reliably across different data partitions. To further assess generalisation performance, the final model was evaluated on the independent hold-out test set. Table 9 reports the evaluation results of the final XGBoost + SMOTE + Boruta model on this unseen dataset.

Table 9. Diabetes Classification Model Performance

Model	Recall	ROC-AUC	PR-AUC
XGBoost + SMOTE + Boruta	0.7500	0.9772	0.8781

Based on Table 9, the XGBoost model combined with SMOTE and Boruta achieved a recall of 0.75 on the hold-out test set, meaning that 75% of diabetic patients were correctly identified. While this

reflects satisfactory sensitivity in detecting the minority class, it also implies that approximately 25% of diabetic cases remained undetected. In absolute terms, out of 1,700 diabetic individuals in the test set, around 425 cases would not be identified by the model. In clinical settings, false negatives are particularly concerning, as undiagnosed diabetes may result in delayed intervention and an increased risk of long-term complications. Therefore, although the recall value indicates reasonable screening capability, the model should be viewed as a decision-support or prioritisation tool rather than a standalone diagnostic system.

The ROC-AUC value of 0.9772 demonstrates strong discriminative performance across classification thresholds, indicating that the model effectively separates diabetic and non-diabetic individuals. A value close to 1 suggests stable ranking ability and consistent class separation. However, in imbalanced datasets, ROC-AUC may yield an overly optimistic assessment because it focuses only on the majority class.

For this reason, PR-AUC is emphasised as a more informative metric in this context. The model achieved a PR-AUC of 0.8781, substantially higher than the baseline value of approximately 0.085, which corresponds to the prevalence of the positive class. This large improvement over baseline indicates that the model maintains a strong balance between precision and recall when identifying diabetic patients. Consequently, the integrated approach combining SMOTE, Boruta, and XGBoost enhances minority class detection while preserving high overall discriminative ability.

3.2.1. Comparison with Baseline Model Without SMOTE

To evaluate the impact of imbalance treatment, a baseline XGBoost model without SMOTE was trained using the same cross-validation scheme. Table 10 presents a comparison of average performance across folds.

Table 10. Comparison of average performance across folds

Metric	SMOTE	No SMOTE	Interpretation
Accuracy	0.9627	0.9717	No SMOTE higher
Precision	0.8085	0.9961	No SMOTE much higher
Recall	0.7396	0.6697	SMOTE higher
F1	0.7716	0.8009	No SMOTE slightly higher
ROC-AUC	0.9757	0.9770	Very similar
MCC	0.7528	0.8043	No SMOTE higher
PR-AUC	0.8720	0.8772	Very similar

A comparison between the SMOTE-enhanced model and the baseline model without imbalance treatment reveals a clear trade-off. Although the baseline model achieved higher precision (0.9961) and slightly higher overall accuracy (0.9717), its recall was substantially lower (0.6697). This indicates that approximately 33% of diabetes cases would remain undetected without imbalance treatment.

After applying SMOTE, recall increased to 0.7396, reducing the proportion of missed diabetes cases from approximately 33% to 26%. In a clinical screening context, false negatives are more critical than false positives because delayed diagnosis may lead to severe complications. Therefore, the improvement in recall enhances the clinical relevance of the proposed model. Although precision decreased after applying SMOTE, the model maintained strong discriminatory capability, as demonstrated by comparable ROC-AUC (0.9757 vs 0.9770) and PR-AUC (0.8720 vs 0.8772) values. This suggests that SMOTE improves minority class detection without significantly degrading overall discrimination performance. The baseline model achieved a higher MCC (0.8043) than the SMOTE model (0.7528),

indicating a stronger overall correlation between predicted and true labels. However, given the clinical priority of minimising false negatives, the improved recall achieved with SMOTE provides a more balanced, clinically meaningful performance.

Therefore, the choice between the two approaches depends on the intended application: the baseline model offers stronger overall performance, whereas the SMOTE-enhanced model provides improved sensitivity to diabetes cases.

Fig. 4 presents a boxplot comparison of cross-validation metrics for models with and without SMOTE, highlighting the consistent improvement in recall after applying imbalance handling.

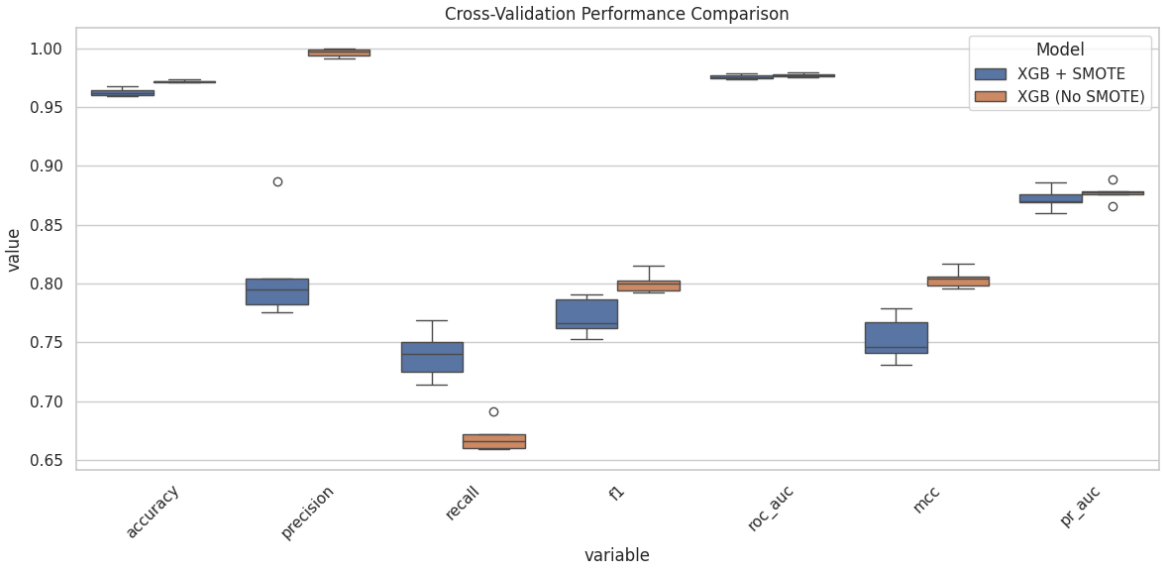


Fig. 4. Boxplot comparison of cross-validation metrics (accuracy, precision, recall, F1-score, ROC-AUC, MCC, and PR-AUC) for XGBoost models with and without SMOTE.

3.2.2. Confusion Matrix Analysis

A confusion matrix provides a detailed overview of the model's predictions on test data, including the counts of true positives, true negatives, false positives, and false negatives, as shown in Fig. 5.

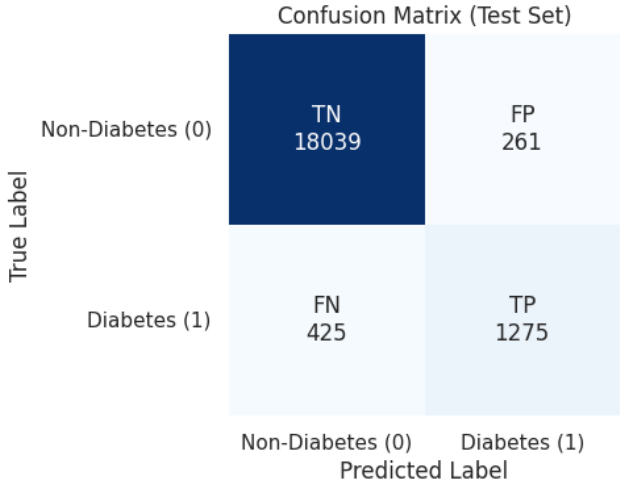


Fig. 5. XGBoost Confusion Matrix on Test Data

The confusion matrix analysis shows that the model achieved high specificity (98.6%), indicating that non-diabetic patients were rarely misclassified. Only 261 out of 18,300 non-diabetic individuals were falsely predicted as diabetic, corresponding to a low false positive rate (1.4%).

From a clinical perspective, false positives may lead to additional laboratory testing or follow-up examinations, which could increase short-term healthcare costs but generally pose minimal medical risk.

In contrast, false negatives represent undiagnosed cases of diabetes. The model produced 425 false negatives out of 1,700 actual diabetes cases, yielding a false-negative rate of 25%. Clinically, false negatives are more concerning because delayed diagnosis may postpone intervention and increase the risk of complications such as cardiovascular disease, nephropathy, neuropathy, and retinopathy.

Although the achieved recall of 0.75 demonstrates strong detection capability for an imbalanced dataset, further threshold optimisation may be considered in future research to prioritise sensitivity depending on screening objectives.

3.2.3. ROC Curve Analysis

The Receiver Operating Characteristic (ROC) curve is used to evaluate a model's ability to distinguish between diabetes and non-diabetes classes at various threshold values, as shown in Fig. 6. This curve visualises the relationship between the True Positive Rate (Recall) and the False Positive Rate, allowing it to be used to assess the overall class separability.

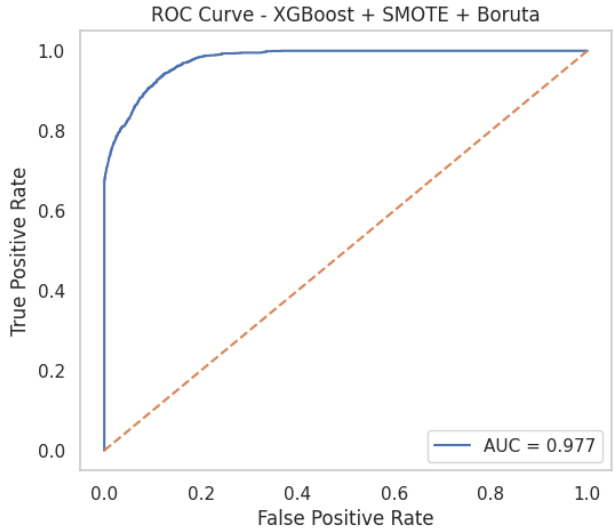


Fig. 6. ROC curve on the hold-out test set showing strong discriminative performance (AUC = 0.9772).

Based on the test data, the ROC curve shows a pattern approaching the top-left corner, with an ROC-AUC of 0.977. This high value indicates that the pipeline used can form a clear decision boundary between the two classes. However, for data with significant class imbalance, ROC-AUC alone is insufficient to characterise the model's performance on the minority class, so additional analysis using the precision-recall curve is needed.

3.2.4. Precision–Recall Curve Analysis

The Precision-Recall (PR) curve provides a more relevant overview in evaluating model performance on imbalanced data because it directly focuses the analysis on the positive class, which is diabetes patients.

This curve shows the relationship between precision and recall at various threshold values, as shown in Fig. 7.

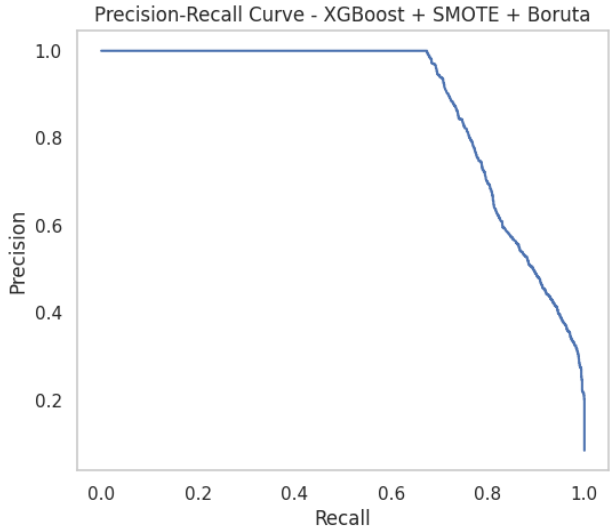


Fig. 7. Precision-Recall Curve

The PR curve results show that the model maintains a relatively high precision value across various recall levels, with a PR-AUC of 0.8781. The curve is well above the baseline line, representing the proportion of positive cases in the dataset. The high PR-AUC value indicates a good balance between precision and recall, and demonstrates that the model can effectively detect diabetic patients without significantly increasing errors.

3.2.5. Cross-Validation and Hold-Out Consistency Analysis

The performance obtained on the hold-out test set closely aligns with the cross-validation results reported in Table 8. For instance, the cross-validation recall (0.7396 ± 0.0192) is consistent with the test recall (0.7500), while ROC-AUC values differ by less than 0.002 (0.9757 vs 0.9772). Similarly, PR-AUC slightly increased from 0.8720 during cross-validation to 0.8781 on the test set.

This close agreement indicates that the proposed SMOTE–Boruta–XGBoost pipeline generalises well to unseen data and does not exhibit signs of overfitting or data leakage. Cross-validation provides an estimate of model stability across different data partitions, whereas hold-out testing offers an unbiased assessment of model performance on completely unseen data. The consistency between these evaluation strategies strengthens confidence in the robustness and reliability of the proposed framework for potential clinical screening applications.

3.3. Model Interpretation Based on SHAP

3.3.1. Identifying the Most Influential Features for Diabetes Prediction

To understand the factors contributing to the diabetes classification model's decisions, an interpretability analysis was conducted using the Shapley Additive Explanations (SHAP) method, as shown in Fig. 8. This approach allows for the quantitative, consistent identification of each feature's contribution to the model's predictions. Global analysis was conducted using SHAP Summary Plot and SHAP Feature Importance (mean absolute SHAP value) to identify the features that contribute the most to overall diabetes predictions.

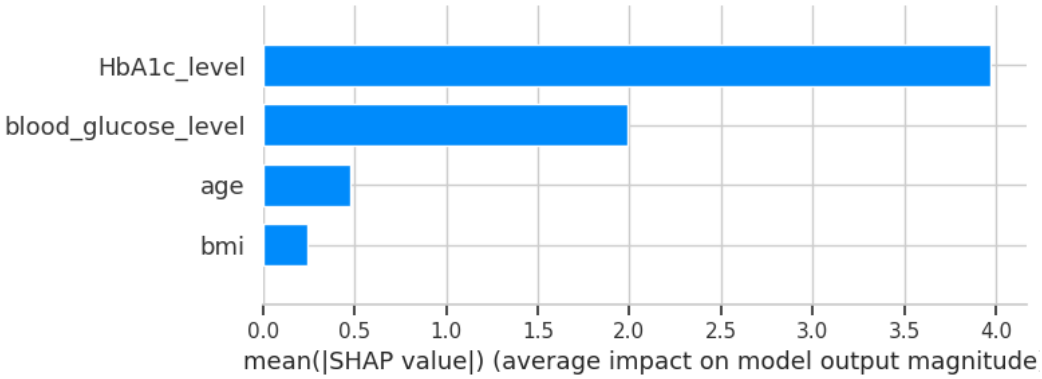


Fig. 8. SHAP Summary Plot (Global Importance / Beeswarm)

The SHAP Summary Plot results show that four main features have a dominant influence on diabetes prediction, namely HbA1c_level, blood_glucose_level, age, and bmi. The order of feature importance, based on mean absolute SHAP values, indicates that HbA1c_level is the most influential factor, followed by blood_glucose_level, age, and bmi. These findings are consistent with the feature importance results from the XGBoost model, showing consistency between the SHAP-based interpretability approach and the model's internal mechanism.

3.3.2. Analysis of Individual Contributions

Further analysis of the direction of feature influence was carried out using SHAP dependence plots. Fig. 9 shows how the values of certain features contribute positively or negatively to the probability of a diabetes prediction.

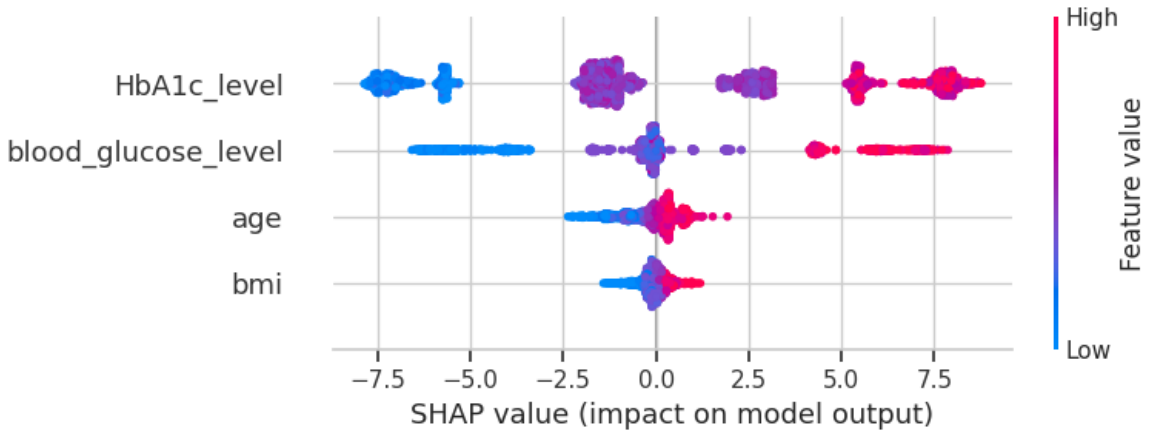


Fig. 9. SHAP Contribution Plot

The contribution plot shows that the model's predictions result from the accumulation of positive and negative contributions from each feature. For patients predicted to have diabetes, the HbA1c_level and blood_glucose_level values generally provide the largest positive contributions to the probability of the diabetes class. Meanwhile, features with values below the clinical threshold can make negative contributions, reducing the predicted probability.

This approach allows for case-based interpretation, enabling medical professionals to understand the specific reasons behind the model's decisions for a particular patient. Thus, the model does not function as a black box but rather as a transparent, clinically auditable system.

3.3.3. Direction of Influence and Nonlinear Relationship of Dominant Features

To understand how changes in feature values affect the probability of a diabetes prediction, an analysis was conducted using a SHAP Dependence Plot on the dominant features.

The HbA1c_level feature, as shown in Fig. 10, makes the largest contribution to diabetes prediction. Higher HbA1c values consistently yield positive SHAP contributions, indicating they increase the model's probability of predicting the patient as diabetic. This reflects the role of HbA1c as a key indicator of long-term glycemic control, which is clinically used to diagnose diabetes [43], [44]. The SHAP value distribution shows that patients with HbA1c above the normal threshold have a much higher risk contribution than those with HbA1c below it. [45] These findings confirm that the model appropriately utilises HbA1c information and aligns with established clinical practice.

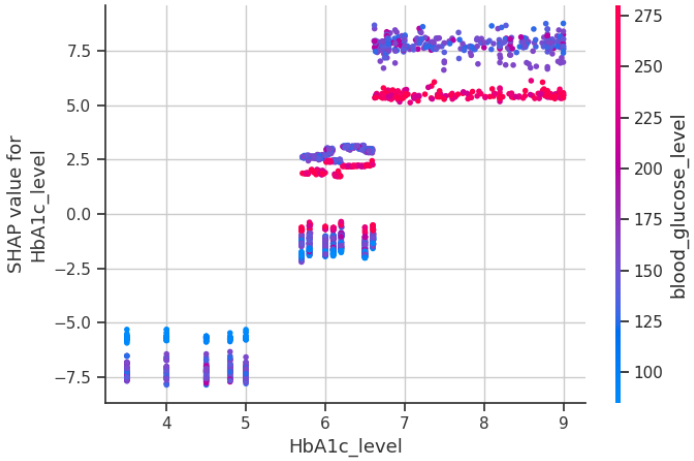


Fig. 10. SHAP Dependence Plot HbA1c_level

The blood_glucose_level feature ranks second in its contribution to diabetes prediction, as shown in Fig. 11. SHAP analysis indicates that higher blood glucose levels are positively correlated with a higher probability of diabetes. High blood glucose values generate significant positive SHAP contributions, indicating the strong role of this feature in the model's decisions [38]. Clinically, blood glucose levels are a direct indicator of a patient's glucose metabolism status [46]. The consistency between SHAP results and medical understanding suggests that the model can recognise relevant physiological relationships rather than merely random statistical patterns.

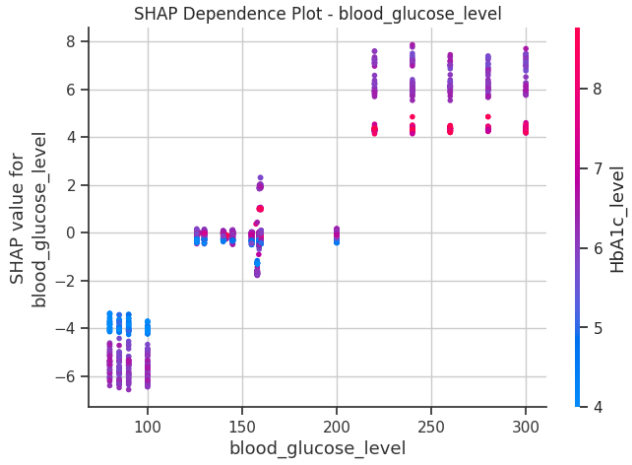


Fig. 11. SHAP Dependence Plot – blood_glucose

Fig. 12 shows a positive association with diabetes prediction, with higher SHAP values for older individuals. This indicates that older age increases the likelihood that a patient will be classified as diabetic by the model [47]. This interpretation aligns with the literature, which states that the risk of diabetes rises with age due to decreased insulin sensitivity and metabolic changes [48]–[50]. Although its contribution is not as significant as that of HbA1c or blood glucose levels, age still consistently influences the model's predictions.

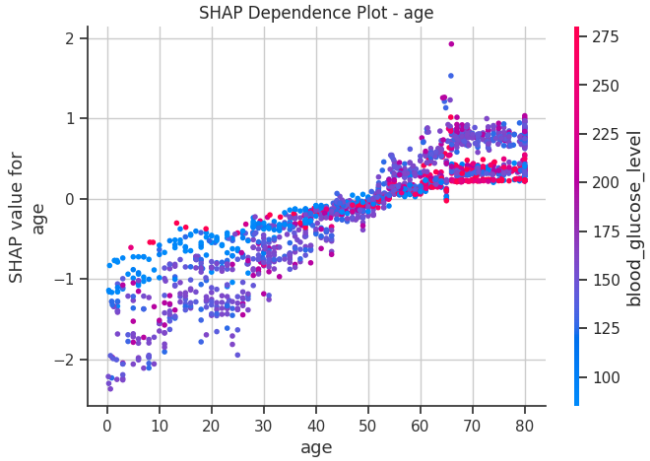


Fig. 12. SHAP Dependence Plot – age

The BMI feature, as shown in Fig. 13, also makes a positive contribution to diabetes prediction, particularly at higher BMI values. SHAP analysis shows that an increase in body mass index raises the probability of diabetes prediction, reflecting the relationship between obesity and insulin resistance [51]–[53]. Although the contribution of BMI is relatively small compared with that of HbA1c and blood glucose levels, this feature still plays an important role in enriching the clinical context of the model's predictions. This indicates that the model does not rely solely on biochemical indicators but also considers medically relevant anthropometric factors.

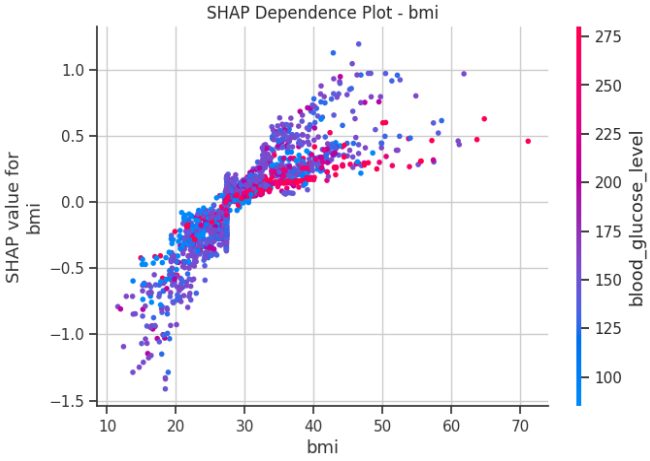


Fig. 13. SHAP Dependence Plot – bmi

SHAP-based interpretation shows that clinically validated diabetes risk factors drive the model's decisions. This indicates that the model is not entirely a black box, but rather produces predictions that can be explained and medically justified.

These findings reinforce that the developed diabetes classification model not only has high predictive performance but also can provide explanations that align with clinical knowledge, making it potentially useful as a clinical decision support system.

3.4. General Discussion and Clinical Implications

The findings demonstrate that integrating class imbalance handling, feature selection, and gradient-boosting-based modelling can yield stable, clinically meaningful predictive performance for diabetes classification. The achieved recall and PR-AUC values suggest that the model can identify a substantial proportion of diabetic individuals while maintaining balanced precision.

From a clinical perspective, however, the implications of misclassification must be carefully considered. A recall of 0.75 indicates that approximately one quarter of diabetic cases remain undetected. In the hold-out test set, this corresponds to approximately 425 missed cases among 1,700 diabetic patients. False negatives are particularly concerning because undiagnosed diabetes may delay treatment initiation and increase the risk of long-term complications such as cardiovascular disease, nephropathy, and neuropathy.

At the same time, the model produces fewer false positives. Although false positives may result in additional laboratory testing or temporary patient anxiety, their clinical consequences are generally less severe than missing true diabetic cases. In large-scale screening contexts, this trade-off is often considered acceptable, as confirmatory testing can resolve uncertain cases with relatively low risk. Therefore, the model should not be viewed as a standalone diagnostic tool but rather as a screening or risk stratification system that supports further confirmatory evaluation.

The strong discriminative performance can be attributed to XGBoost's ability to effectively capture nonlinear relationships and feature interactions, which are commonly present in metabolic disorders. Its built-in regularisation and robustness to correlated clinical variables likely contributed to the stable performance observed across cross-validation folds.

The integration of SHAP-based interpretability further enhances the model's clinical relevance by enabling transparent explanations of prediction drivers. This transparency is essential for building trust among healthcare professionals and ensuring that algorithmic recommendations remain aligned with established medical knowledge and decision-making processes.

Overall, the proposed framework demonstrates potential as a decision-support system to assist early diabetes risk identification, particularly in large-scale population screening scenarios where efficient resource allocation is critical.

4. Conclusion

This study presents an integrated and clinically interpretable machine learning pipeline for diabetes classification that combines imbalance handling (SMOTE), robust feature selection (Boruta), and gradient boosting (XGBoost) within a unified framework. By emphasising evaluation metrics suitable for imbalanced medical data, particularly recall and PR-AUC, the proposed approach demonstrates its effectiveness for minority-class detection while maintaining strong discriminative performance. The integration of SHAP further strengthens the framework by ensuring transparent and clinically meaningful model explanations aligned with established diabetes risk factors. Despite these contributions, several limitations should be acknowledged. The dataset was derived from a single source

and represents static, cross-sectional data, limiting generalizability across diverse populations. External validation on independent cohorts was not conducted, and the model has not yet been evaluated in real-world clinical settings. In addition, decision thresholds were not optimised for clinical cost considerations, which may affect the balance between false negatives and false positives in practical applications. Future research should focus on multi-centre and longitudinal datasets, prospective external validation, and threshold optimisation tailored to specific clinical objectives. Investigating cost-sensitive learning strategies and real-world integration into hospital-based clinical decision support systems would further enhance the translational potential of the proposed framework.

Declarations

Author contribution. All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.

Funding statement. None of the authors has received any funding or grants from any institution or funding body for the research.

Conflict of interest. The authors declare no conflict of interest.

Additional information. No additional information is available for this paper.

References

[1] P. Rajendra and S. Latifi, "Prediction of diabetes using logistic regression and ensemble techniques," *Comput. Methods Programs Biomed. Updat.*, vol. 1, p. 100032, Jan. 2021, doi: 10.1016/j.cmpbup.2021.100032.

[2] C. Azad, B. Bhushan, R. Sharma, A. Shankar, K. K. Singh, and A. Khamparia, "Prediction model using SMOTE, genetic algorithm and decision tree (PMSGD) for classification of diabetes mellitus," *Multimed. Syst.*, vol. 28, no. 4, pp. 1289–1307, Aug. 2022, doi: 10.1007/s00530-021-00817-2.

[3] B. A. C. Permana, R. Ahmad, H. Bahtiar, A. Sudianto, and I. Gunawan, "Classification of diabetes disease using decision tree algorithm (C4.5)," *J. Phys. Conf. Ser.*, vol. 1869, no. 1, p. 012082, Apr. 2021, doi: 10.1088/1742-6596/1869/1/012082.

[4] T. Dudkina, I. Meniailov, K. Bazilevych, S. Krivtsov, and A. Tkachenko, "Classification and Prediction of Diabetes Disease using Decision Tree Method," in *IT&AS 2021: Symposium on Information Technologies & Applied Sciences*, 2021, pp. 1–10. [Online]. Available at: https://ceur-ws.org/Vol-2824/paper16.pdf.

[5] X. Wang *et al.*, "Exploratory study on classification of diabetes mellitus through a combined Random Forest Classifier," *BMC Med. Inform. Decis. Mak.*, vol. 21, no. 1, p. 105, Dec. 2021, doi: 10.1186/s12911-021-01471-4.

[6] H. Hairani, A. Anggrawan, and D. Priyanto, "Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link," *JOIV Int. J. Informatics Vis.*, vol. 7, no. 1, p. 258, Feb. 2023, doi: 10.30630/joiv.7.1.1069.

[7] C. N. Noviyanti and A. Alamsyah, "Early Detection of Diabetes Using Random Forest Algorithm," *J. Inf. Syst. Explor. Res.*, vol. 2, no. 1, pp. 41–48, Jan. 2024, doi: 10.52465/joiser.v2i1.245.

[8] K. Dayal, M. Shukla, and S. Mahapatra, "Disease Prediction Using a Modified Multi-Layer Perceptron Algorithm in Diabetes," *EAI Endorsed Trans. Pervasive Heal. Technol.*, vol. 9, pp. 1–8, Sep. 2023, doi: 10.4108/eetpht.9.3926.

[9] A. A. Safar, D. M. Salih, and A. M. Murshid, "Pattern recognition using the multi-layer perceptron (MLP) for medical disease: A survey," *Int. J. Nonlinear Anal. Appl.*, vol. 14, no. 1, pp. 1989–1998, Jan. 2023. [Online]. Available at: https://ijnaa.semnan.ac.ir/article_7114.html.

[10] A. A. Abu-Shareha, M. Abualhaj, A. Hussein, A. Al-Saaidah, and A. Achuthan, "Investigation of Data Balancing Techniques for Diabetes Prediction," *Int. J. Intell. Eng. Syst.*, vol. 18, no. 3, pp. 598–611, Apr. 2025, doi: 10.22266/ijies2025.0430.41.

[11] N. Sakran *et al.*, "The many faces of diabetes. Is there a need for re-classification? A narrative review," *BMC Endocr. Disord.*, vol. 22, no. 1, p. 9, Jan. 2022, doi: 10.1186/s12902-021-00927-y.

[12] Z. Rafie, M. S. Talab, B. E. Z. Koor, A. Garavand, C. Salehnasab, and M. Ghaderzadeh, "Leveraging XGBoost and explainable AI for accurate prediction of type 2 diabetes," *BMC Public Health*, vol. 25, no. 1, p. 3688, Oct. 2025, doi: 10.1186/s12889-025-24953-w.

[13] W. Li, Y. Peng, and K. Peng, "Diabetes prediction model based on GA-XGBoost and stacking ensemble algorithm," *PLoS One*, vol. 19, no. 9, p. e0311222, Sep. 2024, doi: 10.1371/journal.pone.0311222.

[14] K. H. Abushahla and M. A. Pala, "Optimizing Diabetes Prediction: Addressing Data Imbalance with Machine Learning Algorithms," *ADBA Comput. Sci.*, vol. 1, no. 1, pp. 26–35, Jul. 2024, doi: 10.69882/adba.cs.2024075.

[15] M. M. Jogo Samodro, M. K. Biddinika, and A. Fadlil, "Optimal Feature Selection in Diabetes Classification Using the MLP Algorithm," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 18, no. 2, Apr. 2024, doi: 10.22146/ijccs.94575.

[16] X. Feng, Y. Cai, and R. Xin, "Optimizing diabetes classification with a machine learning-based framework," *BMC Bioinformatics*, vol. 24, no. 1, p. 428, Nov. 2023, doi: 10.1186/s12859-023-05467-x.

[17] G. Abdurrahman, H. Oktavianto, and M. Sintawati, "Optimasi Algoritma XGBoost Classifier Menggunakan Hyperparameter Gridesearch dan Random Search Pada Klasifikasi Penyakit Diabetes," *INFORMAL Informatics J.*, vol. 7, no. 3, p. 193, Dec. 2022, doi: 10.19184/isj.v7i3.35441.

[18] C. Hardiyanti P, "Optimizing breast cancer classification using SMOTE, Boruta, and XGBoost," *Sci. Inf. Technol. Lett.*, vol. 6, no. 1, pp. 16–33, May 2025, doi: 10.31763/sitech.v6i1.2109.

[19] B. L. Ortiz *et al.*, "Data Preprocessing Techniques for AI and Machine Learning Readiness: Scoping Review of Wearable Sensor Data in Cancer Care," *JMIR mHealth uHealth*, vol. 12, no. 1, p. e59587, Sep. 2024, doi: 10.2196/59587.

[20] K. Mallikharjuna Rao, G. Saikrishna, and K. Supriya, "Data preprocessing techniques: emergence and selection towards machine learning models - a practical review using HPA dataset," *Multimed. Tools Appl.*, vol. 82, no. 24, pp. 37177–37196, Oct. 2023, doi: 10.1007/s11042-023-15087-5.

[21] A. Palanivinayagam and R. Damaševičius, "Effective Handling of Missing Values in Datasets for Classification Using Machine Learning Methods," *Information*, vol. 14, no. 2, p. 92, Feb. 2023, doi: 10.3390/info14020092.

[22] M. Hossain, A. Devnath, and P. Karmokar, "Handling Missing Values and Outliers in Advanced Data Pre-processing: An Enhancement of Diabetes Classification Accuracy." p. 13, Sep. 25, 2023, doi: 10.21203/rs.3.rs-3364064/v1.

[23] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[24] D. Elreedy, A. F. Atiya, and F. Kamalov, "A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning," *Mach. Learn.*, vol. 113, no. 7, pp. 4903–4923, Jul. 2024, doi: 10.1007/s10994-022-06296-4.

[25] D. Elreedy *et al.*, "A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning," *Mach. Learn. 2023 1137*, vol. 113, no. 7, pp. 4903–4923, Jan. 2023, doi: 10.1007/s10994-022-06296-4.

[26] S. Subbiah, K. S. M. Anbananthen, S. Thangaraj, S. Kannan, and D. Chelliah, "Intrusion detection technique in wireless sensor network using grid search random forest with Boruta feature selection algorithm," *J. Commun. Networks*, vol. 24, no. 2, pp. 264–273, Apr. 2022, doi: 10.23919/JCN.2022.000002.

[27] M. B. Kursa, "Robustness of Random Forest-based gene selection methods," *BMC Bioinformatics*, vol. 15, no. 1, p. 8, Dec. 2014, doi: 10.1186/1471-2105-15-8.

[28] R. Iranzad and X. Liu, "A review of random forest-based feature selection methods for data science education and applications," *Int. J. Data Sci. Anal.*, pp. 1–15, Feb. 2024, doi: 10.1007/s41060-024-00509-w.

[29] M. Galih Pradana, K. Palilingan, Y. Vanli Akay, D. Puspasari Wijaya, and P. Hari Saputro, "Comparison of Multi Layer Perceptron, Random Forest & Logistic Regression on Students Performance Test," in *2022 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, Nov. 2022, pp. 462–466, doi: 10.1109/ICIMCIS56303.2022.10017501.

[30] G. Fisher, "Pretraining Diversity and Clinical Metric Optimization Achieve State-of-the-Art Performance on ChestX-ray14," *medRxiv*. Cold Spring Harbor Laboratory Press, p. 2025.10.25.25338784, Oct. 27, 2025, doi: 10.1101/2025.10.25.25338784.

[31] J. Cook and V. Ramadas, "When to consult precision-recall curves," *Stata J. Promot. Commun. Stat. Stata*, vol. 20, no. 1, pp. 131–148, Mar. 2020, doi: 10.1177/1536867X20909693.

[32] M. Elghobashy, R. Gama, and R. A. Sulaiman, "Investigation and Causes of Spontaneous (Non-Diabetic) Hypoglycaemia in Adults: Pitfalls to Avoid," *Diagnostics*, vol. 13, no. 20, p. 3275, Oct. 2023, doi: 10.3390/diagnostics13203275.

[33] S. L. Cichosz, C. Bender, and O. Hejlesen, "Explainable Machine Learning-Based Approach to Identify People at Risk of Diabetes Using Physical Activity Monitoring," *BioMedInformatics*, vol. 5, no. 1, p. 1, Dec. 2024, doi: 10.3390/biomedinformatics5010001.

[34] R. Diallo, C. Edalo, and O. O. Awe, "Machine Learning Evaluation of Imbalanced Health Data: A Comparative Analysis of Balanced Accuracy, MCC, and F1 Score," in *STEAM-H: Science, Technology, Engineering, Agriculture, Mathematics and Health*, vol. Part F4005, Springer, Cham, 2025, pp. 283–312, doi: 10.1007/978-3-031-72215-8_12.

[35] L. Zhang and L. Jiang, "Game-theoretic SHAP-driven interpretable forecasting of air cargo demand using Bayesian-optimized random forests," *Front. Phys.*, vol. 13, p. 1705687, Oct. 2025, doi: 10.3389/fphy.2025.1705687.

[36] L. Wu, "A review of the transition from Shapley values and SHAP values to RGE," *Statistics (Ber).*, vol. 59, no. 5, pp. 1161–1183, Sep. 2025, doi: 10.1080/02331888.2025.2487853.

[37] R. Wahyudi, A. A. Dawai, D. Stiawan, A. Pranolo, Y. Mao, and A. H. Bagdade, "SHAP-based interpretable deep learning framework for phishing website detection," *Sci. Inf. Technol. Lett.*, vol. 6, no. 2, pp. 66–88, 2025. [Online]. Available at: https://pubs2.ascee.org/index.php/sitech/article/view/2353.

[38] F. Prendin, J. Pavan, G. Cappon, S. Del Favero, G. Sparacino, and A. Facchinetti, "The importance of interpreting machine learning models for blood glucose prediction in diabetes: an analysis using SHAP," *Sci. Rep.*, vol. 13, no. 1, p. 16865, Oct. 2023, doi: 10.1038/s41598-023-44155-x.

[39] A. V. Ponce-Bobadilla, V. Schmitt, C. S. Maier, S. Mensing, and S. Stodtmann, "Practical guide to <scp>SHAP</scp> analysis: Explaining supervised machine learning model predictions in drug development," *Clin. Transl. Sci.*, vol. 17, no. 11, p. e70056, Nov. 2024, doi: 10.1111/cts.70056.

[40] M. J. Raihan, M. A.-M. Khan, S.-H. Kee, and A.-A. Nahid, "Detection of the chronic kidney disease using XGBoost classifier and explaining the influence of the attributes on the model using SHAP," *Sci. Rep.*, vol. 13, no. 1, p. 6263, Apr. 2023, doi: 10.1038/s41598-023-33525-0.

[41] S. M. Lundberg *et al.*, "From local explanations to global understanding with explainable AI for trees," *Nat. Mach. Intell.*, vol. 2, no. 1, pp. 56–67, Jan. 2020, doi: 10.1038/s42256-019-0138-9.

[42] J. H. Joloudari, A. Marefat, M. A. Nematollahi, S. S. Oyelere, and S. Hussain, "Effective Class-Imbalance Learning Based on SMOTE and Convolutional Neural Networks," *Appl. Sci.*, vol. 13, no. 6, p. 4006, Mar. 2023, doi: 10.3390/app13064006.

[43] C. M. Parrinello and E. Selvin, "Beyond HbA1c and Glucose: the Role of Nontraditional Glycemic Markers in Diabetes Diagnosis, Prognosis, and Management," *Curr. Diab. Rep.*, vol. 14, no. 11, p. 548, Nov. 2014, doi: 10.1007/s11892-014-0548-3.

[44] G. Kaiafa *et al.*, "Is HbA1c an ideal biomarker of well-controlled diabetes?," *Postgrad. Med. J.*, vol. 97, no. 1148, pp. 380–383, Jun. 2021, doi: 10.1136/postgradmedj-2020-138756.

[45] D. Tang *et al.*, "Exploration and analysis of risk factors for coronary artery disease with type 2 diabetes based on SHAP explainable machine learning algorithm," *Sci. Rep.*, vol. 15, no. 1, p. 29521, Aug. 2025, doi: 10.1038/s41598-025-11142-3.

[46] I. Stranders *et al.*, "Admission Blood Glucose Level as Risk Indicator of Death After Myocardial Infarction in Patients With and Without Diabetes Mellitus," *Arch. Intern. Med.*, vol. 164, no. 9, p. 982, May 2004, doi: 10.1001/archinte.164.9.982.

[47] L. Lama *et al.*, "Machine learning for prediction of diabetes risk in middle-aged Swedish people," *Heliyon*, vol. 7, no. 7, p. e07419, Jul. 2021, doi: 10.1016/j.heliyon.2021.e07419.

[48] C. W. Chia, J. M. Egan, and L. Ferrucci, "Age-Related Changes in Glucose Metabolism, Hyperglycemia, and Cardiovascular Risk," *Circ. Res.*, vol. 123, no. 7, pp. 886–904, Sep. 2018, doi: 10.1161/CIRCRESAHA.118.312806.

[49] J. Shou, P.-J. Chen, and W.-H. Xiao, "Mechanism of increased risk of insulin resistance in aging skeletal muscle," *Diabetol. Metab. Syndr.*, vol. 12, no. 1, p. 14, Dec. 2020, doi: 10.1186/s13098-020-0523-x.

[50] L.-Y. Huang, C.-H. Liu, F.-Y. Chen, C.-H. Kuo, P. Pitrone, and J.-S. Liu, "Aging Affects Insulin Resistance, Insulin Secretion, and Glucose Effectiveness in Subjects with Normal Blood Glucose and Body Weight," *Diagnostics*, vol. 13, no. 13, p. 2158, Jun. 2023, doi: 10.3390/diagnostics13132158.

[51] W. Lin, S. Shi, H. Huang, J. Wen, and G. Chen, "Predicting risk of obesity in overweight adults using interpretable machine learning algorithms," *Front. Endocrinol. (Lausanne).*, vol. 14, p. 1292167, Nov. 2023, doi: 10.3389/fendo.2023.1292167.

[52] S. Park, C. Kim, and X. Wu, "Development and Validation of an Insulin Resistance Predicting Model Using a Machine-Learning Approach in a Population-Based Cohort in Korea," *Diagnostics*, vol. 12, no. 1, p. 212, Jan. 2022, doi: 10.3390/diagnostics12010212.

[53] J. Hao *et al.*, "TyG-ABSI as a novel metabolic obesity indicator for carotid plaque: an explainable machine learning study using SHAP in low-income population," *BMC Endocr. Disord.*, vol. 25, no. 1, p. 281, Dec. 2025, doi: 10.1186/s12902-025-02099-5.