Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.3, June 2026, pp. 2357~2368
e-ISSN: 2548-6861

Comprehensive Diabetes Risk Prediction Using BRFSS Data: Performance, Explainability, Fairness, and Calibration

Virzan Pasa Nugraha [1]*, Agung Febrian [2]**
* Informatika, Universitas Sebelas April Sumedang
220660121054@student.unsap.ac.id [1], 220660121086@student.unsap.ac.id [2]

Article Info

Article history:

Received 2026-04-09
Revised 2026-05-01
Accepted 2026-05-25

Keyword:

CatBoost,
Diabetes Prediction,
Explainable AI,
Fairness,
Machine Learning.

ABSTRACT

This study aims to develop and evaluate machine learning models for diabetes risk prediction using a comprehensive approach that considers performance, interpretability, fairness, and calibration aspects. The research employs several classification algorithms, including Logistic Regression, Random Forest, XGBoost, and CatBoost, using the BRFSS dataset. The models are evaluated using multiple metrics, including Accuracy, Balanced Accuracy, Precision, Recall, F1-Score, ROC-AUC, Precision-Recall AUC (PR-AUC), Matthews Correlation Coefficient (MCC), and Brier Score. Explainability analysis is conducted using SHAP to understand feature contributions, while fairness and calibration analyses are performed to assess model reliability and bias across demographic groups. The results show that CatBoost achieves the best overall performance, with the highest ROC-AUC and Recall, as well as the lowest Brier Score, indicating better predictive capability and calibration. Explainability analysis reveals that GenHlth, BMI, and Age are the most influential features, while fairness analysis indicates potential disparities across certain age groups. Furthermore, ablation and misclassification analyses highlight key features and areas for model improvement. Overall, this study demonstrates that integrating performance evaluation with explainability and fairness analysis can produce more reliable and interpretable predictive models for healthcare applications.

I. INTRODUCTION

The rapid development of information and communication technology has driven the adoption of machine learning in various fields, especially healthcare [1]. Previous studies have shown that machine learning algorithms such as Logistic Regression, Random Forest, and XGBoost can achieve strong performance in diabetes prediction tasks [2]. In addition, explainable AI techniques such as SHAP and LIME have been increasingly applied to improve the transparency of machine learning predictions in healthcare [3], [4].

In the healthcare domain, machine learning has been widely applied to predict disease risks based on clinical data [1], [5]. Logistic Regression is one of the most commonly used methods due to its interpretability and competitive performance compared to other models in disease prediction tasks [5], [6]. Furthermore, machine learning enables the identification of complex patterns in healthcare data that are difficult to analyze using conventional methods, thereby supporting more accurate decision-making.

As research progresses, ensemble and boosting methods such as Random Forest, XGBoost, and CatBoost have gained increasing attention due to their ability to significantly improve model performance [7], [8], [9]. XGBoost has been shown to effectively handle large-scale and imbalanced datasets, providing stable performance in evaluation metrics such as AUC and Recall [8]. Meanwhile, CatBoost, as a boosting-based algorithm, has demonstrated superior performance compared to other boosting methods in several classification tasks, making it a promising approach for handling complex data [9].

In addition to model performance, evaluation metrics are essential in assessing the effectiveness of machine learning models. Several studies emphasize the importance of using

multiple evaluation metrics, such as accuracy, precision, Recall, and AUC, to obtain a comprehensive understanding of model performance [5], [10]. This is particularly important in healthcare applications, where prediction errors may have significant consequences for decision-making. Moreover, recent studies highlight the importance of explainable AI techniques, such as SHAP and LIME, to enhance user trust and provide transparent reasoning behind machine learning predictions in healthcare applications [3], [4].

However, most existing studies primarily focus on improving classification performance without considering other critical aspects such as fairness, calibration, and demographic bias across different population groups [1], [5], [10], [11]. In real-world healthcare applications, machine learning models must not only be accurate but also fair across different demographic groups and capable of producing reliable probability estimates through proper calibration analysis [11], [10]. In addition, advanced evaluation techniques and explainable AI approaches are necessary to ensure model stability, transparency, and generalizability across different data conditions [4], [10], [12].

Therefore, this study aims to develop and evaluate machine learning models for diabetes risk prediction using a more comprehensive approach by considering performance, interpretability, fairness, and calibration aspects. This research is expected to contribute to the development of predictive models that are not only accurate but also interpretable and potentially reliable for healthcare applications.

This study makes several key contributions. First, it integrates performance evaluation, explainability, fairness, and calibration within a unified experimental framework. Second, it applies nested stratified cross-validation combined with statistical significance testing to ensure robust model comparison. Third, it incorporates additional analyses including ablation study, misclassification analysis, and threshold optimization to provide deeper insights into model behavior. Unlike previous studies that focus primarily on predictive performance, this research emphasizes model reliability, interpretability, and fairness in healthcare applications.

II. METHOD

A. Research Stages

This study is conducted through a series of systematic stages, including data processing, model development, and evaluation. In general, the research workflow begins with dataset preparation, followed by data preprocessing, data splitting, model development using several machine learning algorithms, and model evaluation. This structured approach aims to ensure that the resulting model achieves optimal performance and can be properly interpreted.

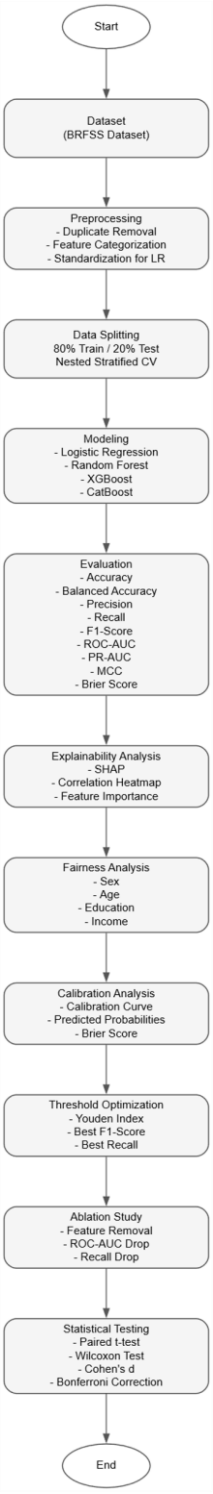

Figure 1. Research Stages

B. Dataset

The dataset used in this study has been selected and prepared according to the requirements of the classification task. The dataset used in this study is derived from the Behavioral Risk Factor Surveillance System (BRFSS) 2015,

a large-scale public health survey conducted by the Centers for Disease Control and Prevention (CDC). The dataset consists of 70,692 survey responses and has been preprocessed into a balanced format with a 50–50 distribution between non-diabetes and diabetes/prediabetes cases. It contains 21 feature variables and one binary target variable (Diabetes_binary), where 0 indicates no diabetes and 1 indicates prediabetes or diabetes.

Previous studies have shown that datasets in machine learning are typically divided into training and testing sets to ensure that the model can generalize well to unseen data [5], [12]. Therefore, in this study, the dataset is further processed before being used in the modeling stage.

The BRFSS dataset initially consisted of 70,692 records and 22 variables. No missing values were found in the dataset; however, 1,635 duplicate rows were identified and removed during preprocessing. After duplicate removal, the final dataset contained 69,057 records with 21 predictor variables and one target variable. The target distribution remained relatively balanced, consisting of 33,960 non-diabetes cases and 35,097 diabetes cases.

The target variable was relatively balanced from the beginning of the study. Before duplicate removal, the dataset contained approximately equal proportions of diabetes and non-diabetes cases. After removing 1,635 duplicate rows, the final dataset consisted of 35,097 diabetes cases (0.5082) and 33,960 non-diabetes cases (0.4918). Since the class distribution remained balanced, no additional resampling techniques such as oversampling, undersampling, or SMOTE were required.

TABLE I
DATASET CHARACTERISTICS

Characteristics	Value
Initial Records	70,692
Duplicate Rows	1,635
Final Records	69,057
Missing Values	0
Mean BMI	29.86
Mean GenHlth	2.84
HighBP Proportion	0.5635
HighChol Proportion	0.5257
Diabetes Cases	35,097
Non-Diabetes Cases	33,960

C. Data Preprocessing

The preprocessing stage is performed to improve data quality before being used in model training. This process includes data cleaning, handling missing values, and transforming the data to fit the requirements of machine learning algorithms [5].

Several studies have highlighted that preprocessing plays a crucial role in improving model performance, particularly in reducing noise and data inconsistencies [13]. Additionally, preprocessing techniques such as normalization, duplicate removal, and feature engineering can significantly enhance classification results [14].

The preprocessing stage included duplicate removal, feature-target separation, and feature categorization into numerical and binary variables. Numerical variables included BMI, GenHlth, MentHlth, PhysHlth, Age, Education, and Income, while binary variables included HighBP, HighChol, CholCheck, Smoker, Stroke, HeartDiseaseorAttack, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, AnyHealthcare, NoDocbcCost, DiffWalk, and Sex. Standardization was applied only to numerical variables in the Logistic Regression model, whereas tree-based models such as Random Forest, XGBoost, and CatBoost were trained without feature scaling.

In addition, multicollinearity analysis was conducted on the numerical variables using Variance Inflation Factor (VIF). All numerical features showed VIF values below 2.0, indicating that no serious multicollinearity issues were present in the dataset. Therefore, all numerical variables were retained for the modeling stage.

After feature-target separation, the predictor matrix consisted of 69,057 rows and 21 input variables, while the target vector contained 69,057 labels.

The predictor variables consisted of 14 binary features and 7 numerical features.

D. Data Splitting

After preprocessing, the dataset is divided into two parts: training data and testing data. This step aims to train the model while also evaluating its performance on unseen data.

Commonly, the dataset is split using a specific ratio, such as 80% for training and 20% for testing, to ensure that the model learns effectively while maintaining objective evaluation [1], [12].

In this study, the dataset was first divided into training and testing sets using an 80:20 ratio with stratification to preserve the class distribution. The training set contained 55,245 records, while the testing set contained 13,812 records. Furthermore, nested stratified cross-validation was applied to ensure robust model evaluation and hyperparameter tuning. The outer loop used 10-fold stratified cross-validation for

performance estimation, while the inner loop used 5-fold stratified cross-validation combined with GridSearchCV for hyperparameter optimization.

E. Modeling

In this stage, classification models are developed using several machine learning algorithms, including Logistic Regression, Random Forest, XGBoost, and CatBoost. Each algorithm has different characteristics in handling classification tasks [5], [7].

Logistic Regression is used due to its capability in handling binary classification and providing probabilistic interpretations [15]. Logistic Regression is used as a baseline model due to its simplicity and interpretability, allowing comparison with more complex models. Random Forest is employed because it improves accuracy by combining multiple decision trees and reducing overfitting [7], [12]. In addition, boosting-based algorithms such as XGBoost and CatBoost are utilized due to their ability to enhance model performance through iterative error optimization.

Hyperparameter optimization was performed using GridSearchCV within the inner cross-validation loop. For Logistic Regression, the optimization focused on the regularization strength parameter (C), while using L2 regularization and the lbfgs solver. For Random Forest, the number of estimators and maximum tree depth were evaluated. Meanwhile, XGBoost and CatBoost were optimized using combinations of learning rate, tree depth, number of estimators or iterations, and regularization-related parameters. This process aimed to identify the best parameter configuration for each model before final evaluation.

TABLE II
HYPERPARAMETER SEARCH SPACE

Model	Hyperparameter
CatBoost	iterations: [100, 200], learning_rate: [0.05, 0.1], depth: [4, 6]
XGBoost	n_estimators: [100, 200], learning_rate: [0.05, 0.1], max_depth: [3, 5]
Random Forest	n_estimators: [100, 200], max_depth: [10, 20], min_samples_split: [2, 4]
Logistic Regression	C: [0.1, 1, 10], penalty: ['l2'], solver: ['lbfgs']

F. Evaluation

The evaluation stage is conducted to measure the performance of the developed models. Several evaluation metrics are used, including Accuracy, Balanced Accuracy, Precision, Recall, F1-Score, ROC-AUC, Precision-Recall AUC (PR-AUC), Matthews Correlation Coefficient (MCC), and Brier Score. In addition, False Positive Rate (FPR) and False Negative Rate (FNR) are analyzed to better understand model errors.

Using multiple evaluation metrics provides a more comprehensive understanding of model performance, especially in classification problems with imbalanced data [10]. Furthermore, a confusion matrix is used to analyze prediction results in more detail [12].

Balanced Accuracy is used to provide a more reliable evaluation when class distribution may vary across subsets of the data. PR-AUC is particularly important in medical classification tasks because it focuses on the positive class (diabetes), which is critical for early detection. Meanwhile, Brier Score is used to evaluate the calibration of predicted probabilities, which is essential in healthcare applications where decision-making depends not only on classification but also on the reliability of probability estimates.

G. Explainability, Fairness, and Calibration

Beyond performance evaluation, this study also considers interpretability and fairness aspects of the model. Explainability is performed using the SHAP (SHapley Additive exPlanations) method to understand the contribution of each feature to the model predictions.

Furthermore, fairness analysis is conducted to ensure that the model does not exhibit bias toward specific groups within the dataset.

This evaluation aligns with the concept of Equal Opportunity, where the model is expected to achieve similar true positive rates across different demographic groups. Finally, calibration analysis is applied to evaluate how well the predicted probabilities reflect actual outcomes, ensuring that the model is not only accurate but also reliable in its probability estimations.

H. Statistical Testing

Statistical significance testing was performed to compare the ROC-AUC performance of different models across the outer cross-validation folds. Paired t-tests and Wilcoxon signed-rank tests were applied to evaluate whether the performance differences between models were statistically significant. In addition, Cohen's d was used to measure effect size, while Bonferroni correction was applied to control for multiple comparisons. This analysis was conducted to ensure that the observed differences in model performance were not caused by random variation alone.

III. RESULTS AND DISCUSSION

A. Model Performance Evaluation

The performance of each classification model was evaluated using multiple metrics, including Accuracy, Balanced Accuracy, Precision, Recall, F1-Score, ROC-AUC,

Precision-Recall AUC (PR-AUC), Matthews Correlation Coefficient (MCC), and Brier Score. These metrics provide a comprehensive assessment of classification performance, model discrimination capability, class balance handling, and probability calibration. The evaluation results are presented in Table III.

TABLE III
MODEL PERFORMANCE COMPARISON

Table III presents the average performance of each model across cross-validation folds. CatBoost achieves the best performance in most evaluation metrics, particularly in ROC-AUC, Recall, and Brier Score.

Model	Accuracy	Balanced Accuracy	Precision	Recall	F1-Score	ROC-AUC	PR-AUC	MCC	Brier Score
CatBoost	0.7495 ± 0.0031	0.7486 ± 0.0031	0.7320 ± 0.0039	**0.8000 ± 0.0047**	0.7645 ± 0.0027	**0.8267 ± 0.0038**	0.8073 ± 0.0065	0.5003 ± 0.0061	**0.1684 ± 0.0018**
XGBoost	0.7494 ± 0.0042	0.7486 ± 0.0043	0.7322 ± 0.0051	0.7994 ± 0.0039	0.7643 ± 0.0035	0.8266 ± 0.0039	0.8073 ± 0.0066	0.5001 ± 0.0083	0.1685 ± 0.0019
Random Forest	0.7468 ± 0.0044	0.7459 ± 0.0044	0.7298 ± 0.0051	0.7969 ± 0.0048	0.7619 ± 0.0038	0.8227 ± 0.0040	0.8033 ± 0.0066	0.4949 ± 0.0087	0.1713 ± 0.0017
Logistic Regression	0.7443 ± 0.0028	0.7439 ± 0.0028	0.7380 ± 0.0035	0.7704 ± 0.0033	0.7539 ± 0.0024	0.8195 ± 0.0034	0.7953 ± 0.0062	0.4886 ± 0.0055	0.1723 ± 0.0017

The results indicate CatBoost achieved the highest average performance across most evaluation metrics among the evaluated models. Although CatBoost showed slightly better results than XGBoost, the difference was very small and was not statistically significant. CatBoost obtained the highest ROC-AUC (0.8267 ± 0.0038), Recall (0.8000 ± 0.0047), and MCC (0.5003 ± 0.0061), while also producing the lowest Brier Score (0.1684 ± 0.0018). These findings indicate that CatBoost provides superior discriminative capability, better sensitivity in identifying positive diabetes cases, and more reliable probability estimates. This performance advantage may be attributed to CatBoost's ordered boosting mechanism, which reduces prediction shift and overfitting during training. Additionally, CatBoost is designed to handle categorical-like patterns and noisy structured data effectively, which aligns well with the characteristics of healthcare datasets such as BRFSS. These properties enable the model to capture complex nonlinear relationships more accurately compared to traditional and bagging-based methods.

Although XGBoost achieved performance values that were very close to CatBoost, CatBoost consistently outperformed the other models across nearly all evaluation metrics. The performance gap between CatBoost and XGBoost was extremely small, with only a 0.0001 difference in ROC-AUC and a 0.0006 difference in Recall. This indicates that both boosting-based models provide similarly strong predictive capability for diabetes risk prediction. Random Forest and Logistic Regression showed relatively lower performance, particularly in terms of ROC-AUC, Recall, and MCC. Logistic Regression achieved the highest Precision (0.7380 ± 0.0035), indicating that it generated fewer false positive predictions; however, its lower Recall suggests that it was less effective in identifying actual diabetes cases.

In healthcare applications, Recall is particularly important because false negative predictions may lead to undetected diabetes cases and delayed treatment. Therefore, CatBoost is considered the most suitable model for diabetes risk prediction due to its strong balance between predictive performance, sensitivity, and calibration.

B. *ROC Curve and PR Curve Analysis*

To further evaluate the classification performance, ROC and Precision-Recall curves were analyzed for all models.

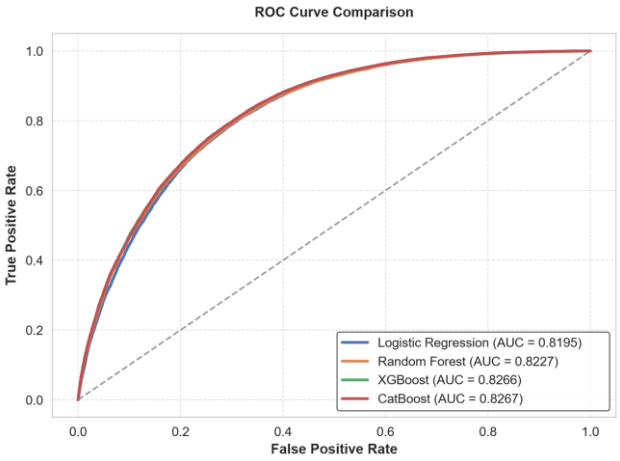

Figure 2. ROC Curve Comparison

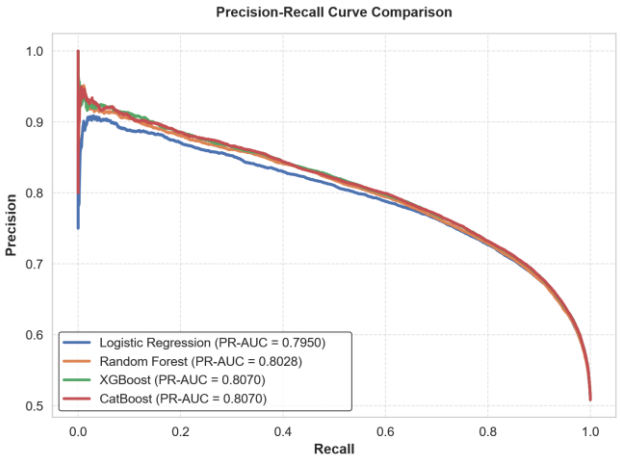

Figure 3. Precision-Recall Curve Comparison

ROC and Precision-Recall curves were analyzed to further evaluate model performance, as shown in Figures 2 and 3. CatBoost achieved the highest ROC-AUC, indicating better discriminative ability. The Precision-Recall curve also shows that CatBoost maintains a better balance between precision and recall, which is important for accurately identifying diabetes cases.

C. Confusion Matrix Analysis

The confusion matrix is used to provide a detailed analysis of prediction outcomes for the best-performing model.

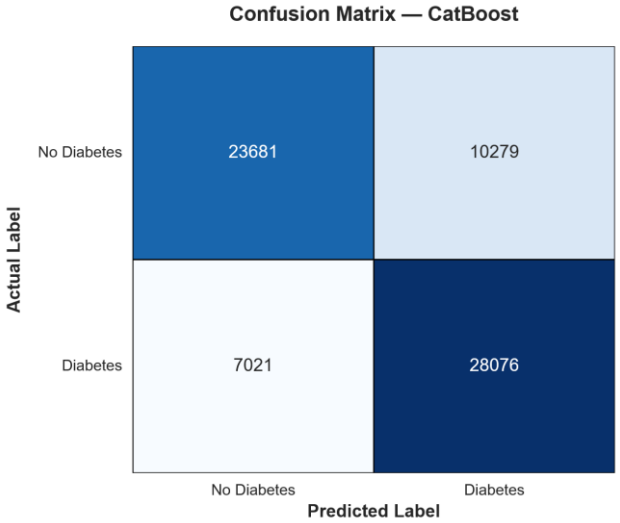

Figure 4. Confusion Matrix of the CatBoost Model

The confusion matrix provides a detailed analysis of prediction outcomes for the best-performing model, as shown in Figure 4. The results show a high number of correct predictions; however, some false negatives still occur. Since false negatives represent missed diabetes cases, a model with higher Recall, such as CatBoost, is more appropriate for healthcare applications.

The confusion matrix of the CatBoost model showed 7,021 false negatives and 10,279 false positives. Although the number of false positives was higher, the relatively lower false negative count is preferable in healthcare applications because missed diabetes cases may lead to delayed diagnosis and treatment. This result indicates that the model is more likely to over-predict diabetes risk rather than fail to identify actual diabetes cases.

The confusion matrix shown in Figure 4 summarizes prediction outcomes aggregated across the outer cross-validation folds. In contrast, the misclassification analysis in Table VI was performed on the final held-out test set. Therefore, the false negative and false positive counts reported in these two sections are not directly comparable because they were obtained from different evaluation settings.

D. Calibration Analysis

Calibration analysis was conducted to evaluate how well the predicted probabilities correspond to the actual outcomes.

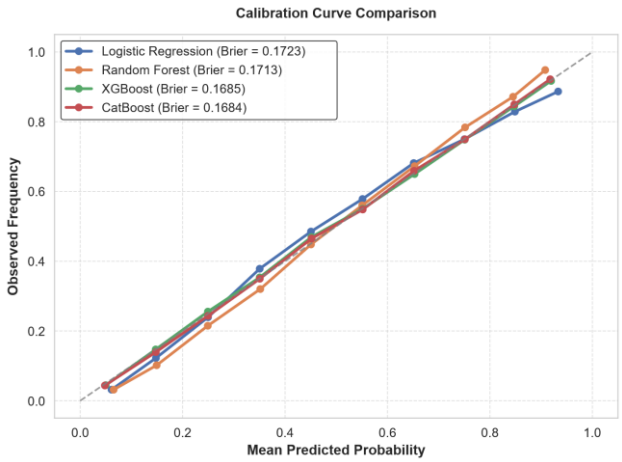

Figure 5. Calibration Curve Comparison

Calibration analysis evaluates how well predicted probabilities match actual outcomes, as shown in Figure 5. CatBoost achieved the lowest Brier Score (0.1684), indicating better calibration and more reliable predicted probabilities, which are essential for real-world healthcare decision-making.

E. Explainability Analysis

Explainability analysis was performed to understand how input features influence model predictions.

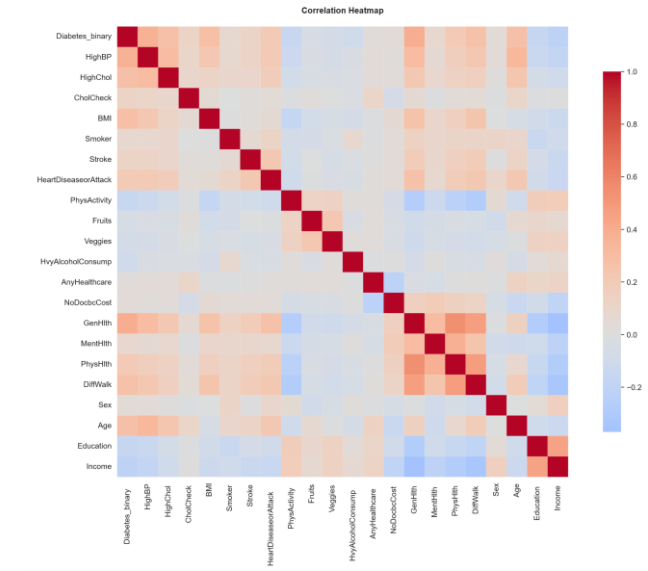

Figure 6. Correlation Heatmap

Figure 6 shows the correlation heatmap among all variables in the dataset. Diabetes status is positively correlated with features such as HighBP, HighChol, BMI, GenHlth, DiffWalk, and Age. In contrast, Education and Income show relatively weak negative correlations with diabetes status. Overall, the correlations are moderate, indicating that no severe multicollinearity exists among the predictor variables.

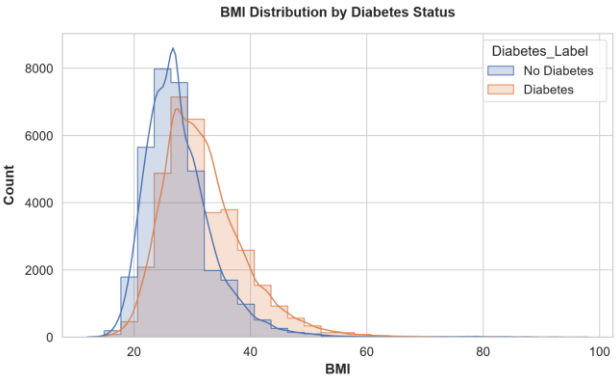

Figure 7. BMI Distribution by Diabetes Status

Figure 7 illustrates the BMI distribution for diabetes and non-diabetes groups. Individuals with diabetes tend to have higher BMI values compared to those without diabetes. The diabetes group shows a distribution concentrated around BMI values above 25, indicating that higher BMI is associated with increased diabetes risk.

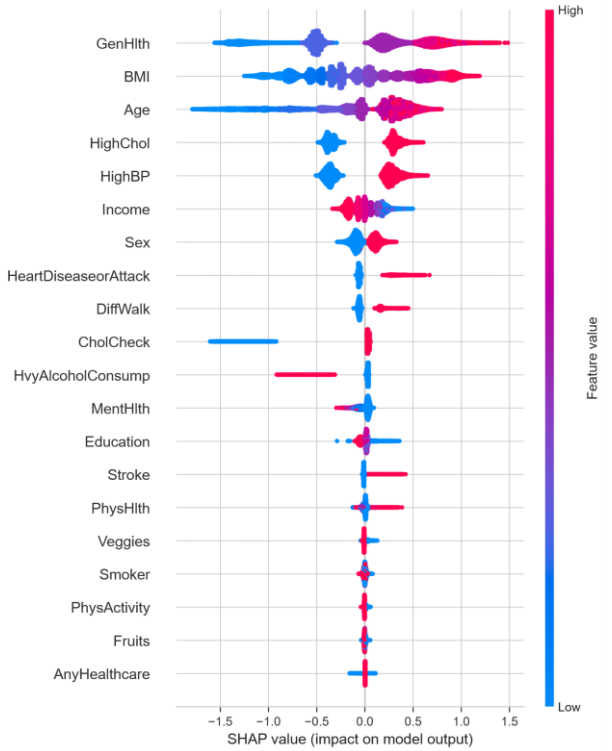

Figure 8. Feature Importance/SHAP Analysis

The SHAP summary plot shows that GenHlth, BMI, and Age are the most influential features in the model. In addition to the summary plot, dependence plots were used to analyze the interaction between key features such as BMI and Age, providing deeper insights into how feature values influence the predicted diabetes risk.

Higher values of these features are associated with an increased probability of diabetes prediction. Additionally, HighBP and HighChol also contribute significantly to the model output.

These results indicate that the model effectively captures key health-related risk patterns, where both general health condition and physiological factors play an important role in diabetes prediction. SHAP enhances model interpretability by providing clear insights into feature contributions, which is essential in healthcare applications.

These findings are consistent with clinical knowledge, where general health condition, obesity (BMI), and aging are well-established risk factors for diabetes. This alignment between model interpretation and domain knowledge increases the trustworthiness of the predictive model.

In terms of mean absolute SHAP values, GenHlth had the highest contribution to the model predictions (0.5586), followed by BMI (0.4721), Age (0.4011), HighChol (0.3380), and HighBP (0.3226). These results further confirm that general health condition, body mass index, age, cholesterol status, and blood pressure are the most influential factors in diabetes risk prediction.

F. Fairness Analysis

Fairness analysis was conducted to assess whether the best-performing model exhibits performance disparities across different demographic groups. Figure 9 presents the age distribution across diabetes and non-diabetes groups. Diabetes cases are more common in older age categories, particularly groups 9 to 13. In contrast, younger age groups are dominated by non-diabetes cases. This pattern indicates that diabetes prevalence increases with age.

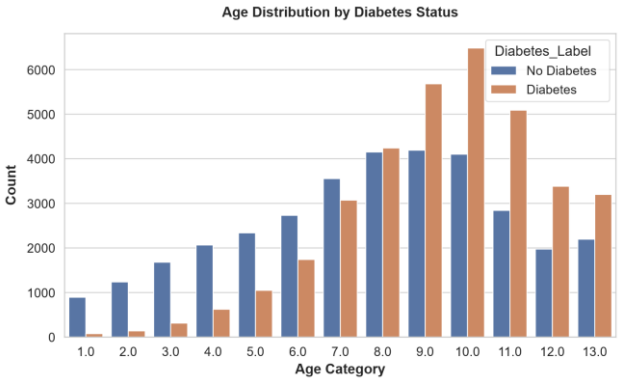

Figure 9. Age Distribution by Diabetes Status

The evaluation was performed based on Sex, Age, and Education using several metrics, including Accuracy, Recall, False Positive Rate (FPR), False Negative Rate (FNR), and ROC-AUC. The results are presented in Table IV.

TABLE IV
FAIRNESS EVALUATION RESULTS

Feature	Group	Sample Size	Accuracy	Recall	FPR	FNR	ROC-AUC
Sex	0	7438	0.7606	0.7986	0.2757	0.2014	0.8375
Sex	1	6374	0.7344	0.8012	0.3415	0.1988	0.8074
Age	1	197	0.9340	0.0769	0.0054	0.9231	0.8537
Age	2	279	0.9319	0.3043	0.0117	0.6957	0.8894
Age	3	383	0.8564	0.3387	0.0436	0.6613	0.8290
Age	4	549	0.8106	0.3561	0.0456	0.6439	0.8535
Age	5	673	0.7816	0.5856	0.1220	0.4144	0.8482
Age	6	882	0.7619	0.6830	0.1869	0.3170	0.8308
Age	7	1324	0.7432	0.7342	0.2493	0.2658	0.8156
Age	8	1631	0.7315	0.7756	0.3132	0.2244	0.8144
Age	9	2015	0.7424	0.8206	0.3610	0.1794	0.7996
Age	10	2142	0.7502	0.8669	0.4435	0.1331	0.7982
Age	11	1568	0.7404	0.8899	0.5094	0.1101	0.7783
Age	12	1060	0.6991	0.8462	0.5599	0.1538	0.7303
Age	13	1109	0.6682	0.8341	0.5730	0.1659	0.7029
Education	2	330	0.7182	0.9144	0.6852	0.0856	0.7680
Education	3	684	0.7617	0.9122	0.5167	0.0878	0.7699
Education	4	3917	0.7302	0.8435	0.4188	0.1565	0.7928
Education	5	3906	0.7611	0.8162	0.2990	0.1838	0.8365
Education	6	4956	0.7534	0.7001	0.2080	0.2999	0.8311

The fairness evaluation shows relatively similar Recall values between the two sex groups, with group 0 achieving a Recall of 0.7986 and group 1 achieving 0.8012. However, group 1 exhibits a higher false positive rate (0.3415) compared to group 0 (0.2757), indicating that the model tends to produce more false positive predictions for this group.

More substantial disparities are observed across age groups. Younger age groups, particularly groups 1 to 4, show very low Recall values ranging from 0.0769 to 0.3561 and very high false negative rates ranging from 0.6439 to 0.9231. In contrast, older age groups, such as groups 10 to 13, achieve much higher Recall values above 0.83 with lower false negative rates. This suggests that the model is more effective in identifying diabetes cases among older individuals than among younger individuals. This disparity may occur because younger age groups contain fewer diabetes-positive cases and smaller sample sizes, causing the model to have limited exposure to diabetes patterns among younger individuals during training.

Differences are also observed across education levels. Lower education groups, especially groups 2 and 3, achieve very high Recall values above 0.91 but also show high false positive rates above 0.50. Meanwhile, higher education groups tend to have lower false positive rates but slightly lower Recall values. These findings indicate that the model sensitivity and error distribution vary across demographic groups.

Overall, the results highlight the importance of fairness evaluation in healthcare prediction models. Although the model demonstrates strong predictive performance, disparities across age and education groups suggest that further bias mitigation strategies may be required to ensure more equitable predictions across the population.

These findings indicate a potential bias toward older age groups, where the model performs significantly better in identifying diabetes cases. In contrast, younger individuals tend to be under-detected, as reflected by lower Recall and higher false negative rates. From a clinical perspective, this may lead to delayed diagnosis in younger populations. From an ethical standpoint, such disparities raise concerns regarding equity in healthcare decision support systems, highlighting the need for bias mitigation strategies in future work.

Fairness analysis was also conducted across income groups. The results show an average Accuracy of 0.7463, Recall of 0.8263, false positive rate of 0.4108, false negative rate of 0.1737, and ROC-AUC of 0.7942 across the analyzed income categories. These findings indicate that model performance also varies across socioeconomic levels, suggesting that income may influence the model's ability to identify diabetes risk consistently across different population groups.

G. Ablation Study

An ablation study was performed to evaluate the impact of individual features on model performance.

TABLE V
ABLATION STUDY RESULTS

Removed Feature	ROC-AUC	Recall	Balanced Accuracy	ROC-AUC Drop	Recall Drop	Balanced Accuracy Drop
BMI	0.8115	0.7992	0.7360	0.0153	0.0008	0.0127
GenHlth	0.8118	0.7841	0.7365	0.0150	0.0159	0.0121
Age	0.8178	0.7854	0.7423	0.0089	0.0145	0.0063
HighBP	0.8203	0.7958	0.7446	0.0065	0.0041	0.0041
HighChol	0.8218	0.7987	0.7460	0.0049	0.0012	0.0026

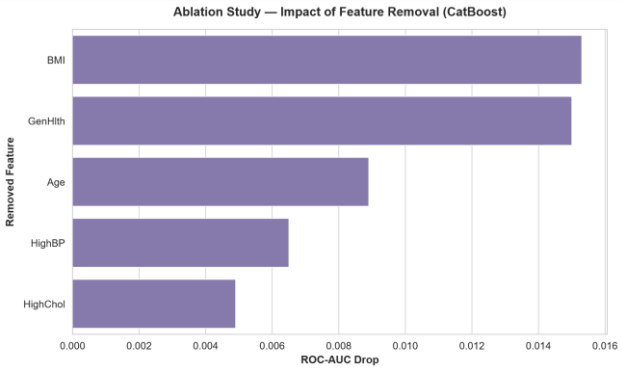

Figure 10. Feature Impact Based on Ablation Study

The results, as presented in Table V and Figure 10, show that removing certain features leads to a decrease in performance, particularly in terms of ROC-AUC. Among all features, BMI caused the largest ROC-AUC reduction (0.0153), while GenHlth caused the largest Recall reduction (0.0159). This indicates that BMI contributes more strongly to overall discrimination performance, whereas GenHlth has a greater impact on identifying positive diabetes cases. In contrast, HighBP and HighChol produced relatively smaller performance reductions, suggesting lower influence on the overall prediction capability of the model.

The ablation results confirm that BMI plays a critical role in model discrimination, as its removal causes the largest decrease in ROC-AUC. Meanwhile, GenHlth has a stronger impact on Recall, indicating its importance in identifying positive diabetes cases.

H. Misclassification Analysis

Misclassification analysis was conducted to further investigate prediction errors, as presented in Table VI and Figure 11.

TABLE VI
MISCLASSIFICATION SUMMARY

Category	Count
False Negative	1405
False Positive	2069
High Confidence Errors	580
Borderline Predictions	1307

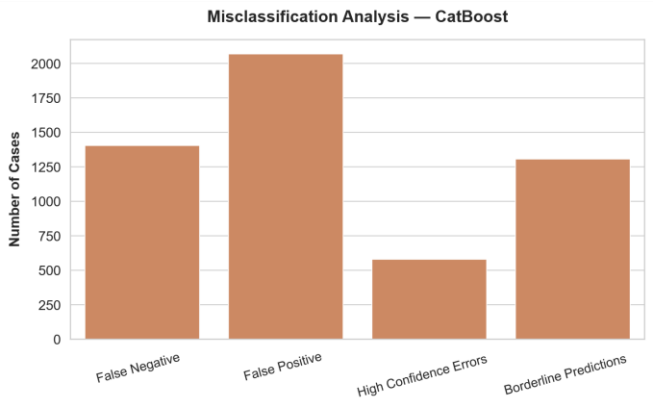

Figure 11. Misclassification Distribution

The results show that the model produces a higher number of false positives (2069) compared to false negatives (1405), indicating a tendency to over-predict positive cases. Additionally, a notable number of borderline predictions (1307) suggests uncertainty in certain cases, while high-confidence errors (580) indicate instances where the model makes incorrect predictions with strong confidence. These findings highlight areas for improvement, particularly in reducing false positives and enhancing prediction reliability.

I. Threshold Optimization Analysis

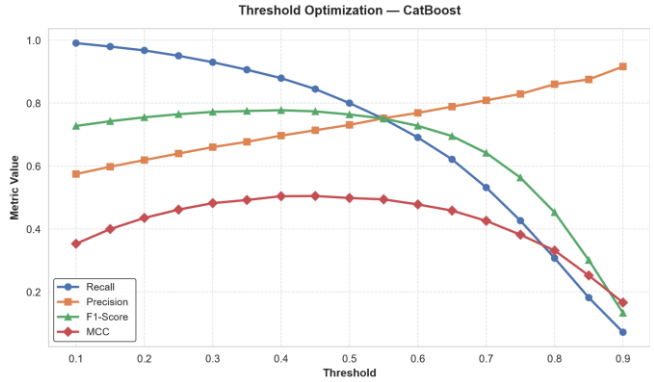

Figure 12. Threshold Optimization Results for CatBoost

Figure 12 illustrates the relationship between Recall, Precision, F1-Score, and MCC under different threshold settings. Lower threshold values tend to increase Recall while decreasing Precision, whereas higher thresholds generally improve Precision but reduce Recall.

Threshold optimization analysis was conducted to evaluate how different probability thresholds affect model performance.

TABLE VII
BEST THRESHOLD SELECTION RESULTS

Criterion	Threshold	Accuracy	Precision	Recall	F1-Score	MCC	False Negatives	False Positives
Best Youden Index	0.50	0.7485	0.7307	0.7999	0.7637	0.4983	1405	2069
Best F1-Score	0.40	0.7440	0.6966	0.8792	0.7773	0.5040	848	2688
Best Recall	0.10	0.6224	0.5745	0.9906	0.7273	0.3529	66	5150

Threshold analysis was conducted on the final test set to determine the optimal classification threshold for the CatBoost model. The default threshold of 0.50 provided the best trade-off between Recall and false positive predictions according to the Youden Index, achieving an Accuracy of 0.7485, Recall of 0.7999, and MCC of 0.4983. At this

threshold, the model produced 1,405 false negatives and 2,069 false positives.

Lowering the threshold to 0.40 increased Recall to 0.8792 and reduced the number of false negatives to 848, indicating that more diabetes cases could be detected earlier. However, this adjustment also increased the number of false positives to 2,688.

Meanwhile, a threshold of 0.10 achieved the highest Recall of 0.9906 but resulted in 5,150 false positives, making it less practical for healthcare screening. In healthcare applications, reducing false negatives is often more important than minimizing false positives because missed diabetes cases may lead to delayed diagnosis and treatment

These results indicate that threshold selection can be adjusted depending on whether minimizing false negatives or false positives is prioritized in healthcare applications. A threshold of 0.40 may be more appropriate for screening purposes because it substantially reduces false negatives while still maintaining relatively balanced overall performance. In contrast, the default threshold of 0.50 remains more suitable for general classification tasks that require a better balance between Recall and false positive predictions. Meanwhile, a threshold of 0.10 may be too sensitive for practical implementation because it produces a very large number of false positive predictions.

J. Multicollinearity Analysis

Multicollinearity analysis was performed using Variance Inflation Factor (VIF) to assess the degree of correlation among numerical features.

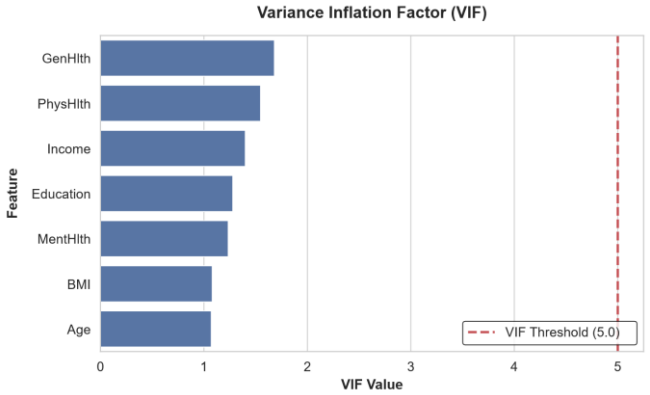

Figure 13. Variance Inflation Factor (VIF) Results

The results show that all numerical variables had VIF values below 2.0, indicating that there were no serious multicollinearity issues in the dataset. The highest VIF values were observed for GenHlth (1.6813) and PhysHlth (1.5474), followed by Income (1.4001), Education (1.2788), and MentHlth (1.2347). Meanwhile, BMI and Age had relatively low VIF values of 1.0802 and 1.0711, respectively. These findings suggest that the numerical features provide complementary information and do not excessively overlap in representing diabetes risk factors.

K. Statistical Significance Testing Results

Statistical significance testing was performed to compare ROC-AUC values between models across the outer cross-validation folds. The results are presented in Table VIII.

TABLE VIII
STATISTICAL SIGNIFICANCE TESTING RESULTS

Model 1	Model 2	Mean Model 1	Mean Model 2	Paired t-test Statistic	Paired t-test p-value	Wilcoxon Statistic	Wilcoxon p-value	Cohen's d	Bonferroni Corrected p-value	Significant Difference
Logistic Regression	Random Forest	0.8195	0.8227	-6.2197	0.0002	0.0	0.0020	-1.9668	0.0009	True
Logistic Regression	XGBoost	0.8195	0.8266	-13.5480	0.0000	0.0	0.0020	-4.2842	0.0000	True
Logistic Regression	CatBoost	0.8195	0.8267	-18.9096	0.0000	0.0	0.0020	-5.9797	0.0000	True
Random Forest	XGBoost	0.8227	0.8266	-15.5321	0.0000	0.0	0.0020	-4.9117	0.0000	True
Random Forest	CatBoost	0.8227	0.8267	-15.4893	0.0000	0.0	0.0020	-4.8981	0.0000	True
XGBoost	CatBoost	0.8266	0.8267	-0.6006	0.5629	20.0	0.4922	-0.1899	1.0000	False

The results show that CatBoost significantly outperformed Logistic Regression and Random Forest after Bonferroni correction. However, the difference between CatBoost and XGBoost was not statistically significant, indicating that both boosting-based models achieved comparable predictive performance. Although CatBoost achieved slightly higher ROC-AUC values than XGBoost, the difference was extremely small, with mean ROC-AUC values of 0.8267 and 0.8266, respectively. The paired t-test produced a p-value of 0.5629, while the Bonferroni-corrected p-value reached 1.0000. These results indicate that the observed performance difference between CatBoost and XGBoost was not statistically significant and may be attributed to random variation across the cross-validation folds.

L. Limitations

One limitation of this study is that the dataset was obtained only from the BRFSS source, which may limit the generalizability of the model to other populations or healthcare settings. Differences in demographic distribution and healthcare systems may limit the model's applicability across regions. In addition, no external validation was performed using independent datasets from different regions or institutions. Although the fairness analysis identified disparities across age, education, and income groups, no bias mitigation techniques were applied in this study. The model also produced a relatively high number of false positive predictions, particularly among older age groups, which may reduce its practicality in certain healthcare scenarios. Furthermore, the model performance was evaluated using a fixed set of features, and additional clinical or laboratory variables may further improve prediction accuracy in future studies. These limitations provide important directions for future research and model refinement.

IV. CONCLUSION

This study presents a comprehensive approach to diabetes risk prediction by evaluating multiple machine learning models alongside explainability, fairness, and calibration analyses. The results show that CatBoost achieves the best overall performance, with superior ROC-AUC, Recall, and calibration, making it more suitable for healthcare applications. Explainability analysis reveals that features such as GenHlth, BMI, and Age are the most influential in model predictions, while fairness analysis highlights potential disparities across certain demographic groups, particularly age. The ablation study further confirms the importance of key features, and misclassification analysis identifies areas for improvement, especially in reducing false positives. These findings extend previous studies by not only focusing on model performance but also incorporating interpretability, fairness, calibration, and statistical validation aspects.

In addition, statistical testing confirmed that CatBoost significantly outperformed Logistic Regression and Random Forest, while showing no statistically significant difference compared to XGBoost. This indicates that both boosting-based models are highly competitive for diabetes risk prediction. Threshold optimization analysis further showed that the model can be adjusted depending on whether minimizing false negatives or false positives is prioritized. This flexibility is particularly important in healthcare applications, where the consequences of missed diagnoses and unnecessary follow-up examinations must be carefully balanced. This model can be used as a decision support tool for early diabetes screening, particularly in identifying high-risk individuals for further clinical examination.

Future work should explore bias mitigation techniques and model optimization to promote more equitable predictions across all population groups.

REFERENCES

[1] A. Pinandito, S. A. Wicaksono, and S. H. Wijoyo, "Implementasi Machine Learning dalam Deteksi Risiko Tinggi Diabetes Melitus pada Kehamilan," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 10, no. 4, pp. 739–746, Aug. 2023, doi: 10.25126/jtiik.2023107005.

[2] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, "Diabetes prediction using machine learning and explainable AI techniques," *Healthc. Technol. Lett.*, vol. 10, no. 1–2, pp. 1–10, Feb. 2023, doi: 10.1049/htl2.12039.

[3] P. Netayawijit, W. Chansanam, and K. Sorn-In, "Interpretable Machine Learning Framework for Diabetes Prediction: Integrating SMOTE Balancing with SHAP Explainability for Clinical Decision Support," *Healthcare*, vol. 13, no. 20, p. 2588, Oct. 2025, doi: 10.3390/healthcare13202588.

[4] R. Alkhanbouli, H. Matar Abdulla Almadhaani, F. Alhosani, and M. C. E. Simsekler, "The role of explainable artificial intelligence in disease prediction: a systematic literature review and future research directions," *BMC Med. Inform. Decis. Mak.*, vol. 25, no. 1, p. 110, Mar. 2025, doi: 10.1186/s12911-025-02944-6.

[5] E. D. C. Pereira and W. Andriyani, "Diabetes Prediction Using Machine Learning," *JIKO (Jurnal Informatika dan Komputer)*, vol. 9, no. 3, p. 639, Oct. 2025, doi: 10.26798/jiko.v9i3.2104.

[6] D. Fabiyanto and Z. Pratama Putra, "Validasi Efektivitas Logistic Regression untuk Diagnosa Penyakit Jantung melalui Pendekatan Machine Learning," *Jurnal Ilmiah FIFO*, vol. 16, no. 2, p. 158, Nov. 2024, doi: 10.22441/fifo.2024.v16i2.006.

[7] I. M. Khoirun Nisa' and R. Nooraeni, "Penerapan Metode Random Forest Untuk Klasifikasi Wanita Usia Subur Di Perdesaan Dalam Menggunakan Internet (SDKI 2017)," *Jurnal MSA (Matematika dan Statistika serta Aplikasinya)*, vol. 8, no. 1, p. 72, Jun. 2020, doi: 10.24252/msa.v8i1.13162.

[8] M. Dzaky and A. Prayogo Kuncoro, "Optimizing XGBoost for Heart Disease Risk Classification Using Optuna and Random Search on the Behavioral Risk Factor Surveillance System (BRFSS) 2023 Dataset," 2026. [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[9] T. Z. Jasman, M. A. Fadhlullah, A. L. Pratama, and R. Rismayani, "Analisis Algoritma Gradient Boosting, Adaboost dan Catboost dalam Klasifikasi Kualitas Air," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 8, no. 2, Aug. 2022, doi: 10.28932/jutisi.v8i2.4906.

[10] B. Van Calster *et al.*, "Evaluation of performance measures in predictive artificial intelligence models to support medical decisions: overview and guidance," *Lancet Digit. Health*, vol. 7, no. 12, Dec. 2025, doi: 10.1016/j.landig.2025.100916.

[11] Q. Feng, M. Du, N. Zou, and X. Hu, "Fair Machine Learning in Healthcare: A Survey," *IEEE Transactions on Artificial Intelligence*, vol. 6, no. 3, pp. 493–507, Mar. 2025, doi: 10.1109/TAI.2024.3361836.

[12] W. O. Simanjuntak, A. Bijaksana, P. Negara, and R. Septriana, "Perbandingan Algoritma Logistic Regression dan Random Forest (Studi Kasus : Klasifikasi Emosi Tweet) Comparison Of Logistic Regression And Random Forest Algorithms (Case Study: Tweet Emotion Classification)," 2023, doi: 10.26418/juara.v2i1.69682.

[13] M. Purba, S. Dianing Asri, V. Ayumi, U. Salamah, and L. Iryani, "Klasifikasi Dataset Teks Pengaduan Masyarakat Terhadap Pemerintah di Sosial Media Menggunakan Logistic Regression," *JSAI (Journal Scientific and Applied Informatics)*, vol. 7, no. 1, pp. 78–83, Jan. 2024, doi: 10.36085/jsai.v7i1.6447.

[14] M. I. Arrasyid Supriyanto, A. A. Rasendrya Hasan, D. Dharmaesa, R. F. Aththar, S. A. Febrinato, and C. M. Sari, "Integrasi Mobile Aplikasi Untuk Klasifikasi Harga Laptop Menggunakan Metode Support Vector Classification Dan Logistic Regression," *Jurnal Media Informatika*, vol. 6, no. 4, pp. 2342–2350, Aug. 2025, doi: 10.55338/jumin.v6i4.6576.

[15] A. Salim and M. R. Alfian, "Optimalisasi Regresi Logistik Menggunakan Algoritma Genetika Pada Data Klasifikasi," *Jurnal Teknologi Informasi dan Terapan*, vol. 6, no. 2, pp. 50–55, Dec. 2019, doi: 10.25047/jtit.v6i2.109.