

Comparative Evaluation of Machine Learning Models with Explainability and Fairness for Diabetes Risk Prediction

Vanya Dwi Nabila¹, Bayu Wijaya Putra^{2*}, M Rudi Sanjaya², Dwi Rosa Indah¹

¹ Faculty of Computer Science, Information System, Universitas Sriwijaya, Palembang, Indonesia

² Faculty of Computer Science, Informatics Management, Universitas Sriwijaya, Palembang, Indonesia

Email: ¹vanyadwinabila@gmail.com, ^{2*}bayuwisata@gmail.com, ³m.rudi.sjy@ilkom.unsri.ac.id, ⁴indah812@unsri.ac.id

Correspondence Author Email: bayuwisata@gmail.com

Received: 18/07/2026, Revised: 04/08/2026, Accepted: 04/09/2026, Published: 08/09/2026

Abstract—Diabetes mellitus remains a major global health challenge, making early risk prediction essential for timely intervention and prevention. This study proposes a trustworthy machine learning framework for diabetes risk prediction by integrating predictive performance, explainability, and fairness evaluation. Six classification algorithms Logistic Regression, Decision Tree, Random Forest, XGBoost, LightGBM, and CatBoost were comparatively evaluated, with LightGBM selected as the best baseline model. Hyperparameter optimization was subsequently performed using Optuna with macro F1-score as the optimization objective to better address the imbalanced multiclass nature of the dataset. Model performance was assessed using Accuracy, Balanced Accuracy, Precision, Recall, F1-score, and Receiver Operating Characteristic–Area Under the Curve (ROC-AUC). Model explainability was analyzed using SHapley Additive exPlanations (SHAP), while fairness was evaluated using Fairlearn based on gender and race through Demographic Parity Difference and Equalized Odds Difference. Experimental results show that hyperparameter optimization increased Balanced Accuracy from 0.3755 to 0.4891 and macro F1-score from 0.3797 to 0.4333, indicating improved recognition of minority classes. Although overall Accuracy decreased from 0.8357 to 0.7023, this trade-off reflects a more balanced classification across diabetes categories, which is preferable for imbalanced clinical datasets where identifying minority cases is essential for early risk detection. SHAP analysis identified Body Mass Index (BMI), Age, Physical Health Days, Mental Health Days, and Income Level as the most influential predictors. Fairness evaluation further demonstrated low demographic disparities across gender and race. These findings demonstrate that integrating predictive performance, explainability, and fairness enables the development of a more transparent, equitable, and clinically reliable diabetes risk prediction framework.

Keywords: Diabetes Risk Prediction; LightGBM; Optuna; Explainable Artificial Intelligence; SHAP; Fairlearn; Fairness; BRFS 2024

1. INTRODUCTION

Among non-communicable diseases, diabetes mellitus continues to rank as one of the most serious, posing an ongoing threat to public health on a global scale. According to figures released by the International Diabetes Federation (IDF), the number of people affected by diabetes around the world continues to rise, with projections pointing to a substantial further increase in the coming years [1]. The disease imposes a heavy toll through illness, mortality, and healthcare spending, primarily due to its strong association with grave complications, including cardiovascular disorders, stroke, chronic kidney disease, neuropathy, and retinopathy. Compounding the problem is the fact that a considerable proportion of affected individuals go undiagnosed until complications have already emerged, which diminishes treatment effectiveness and pushes healthcare costs higher. Given these circumstances, identifying people at risk of diabetes before onset has become a vital strategy, both for preventing the disease itself and for informing clinical practice and evidence-based health policy.

The swift expansion of digital healthcare data has driven greater interest in applying machine learning methods to predict diabetes risk. In contrast to conventional statistical models, machine learning algorithms are able to detect intricate nonlinear patterns linking demographic factors, health status, lifestyle habits, and medical history, which makes them well suited for early disease detection and for supporting clinical decision-making systems [2]. However, improving predictive performance alone is insufficient for healthcare applications because prediction models are also expected to provide transparent and trustworthy explanations that can be understood by healthcare professionals.

A considerable number of studies have built diabetes prediction models employing a range of machine learning algorithms. Das et al. [3] examined several ensemble learning approaches and found that boosting-based methods consistently delivered better performance than conventional classification techniques. Comparable results were also documented by Li et al. [4] and Rafie et al. [5], who demonstrated that XGBoost achieved superior predictive performance on large-scale healthcare datasets. Furthermore, Pang [6] showed that gradient boosting models exhibit strong generalization capability when combined with model interpretability techniques. These findings indicate that ensemble learning has become one of the most promising approaches for diabetes risk prediction.

Beyond algorithm selection itself, hyperparameter optimization has become a critical factor influencing predictive performance in machine learning models. Recent studies have demonstrated that Optuna provides an efficient Bayesian optimization framework for automatically identifying optimal hyperparameter configurations, often outperforming conventional approaches such as Grid Search and Random Search while reducing computational cost [7]. Consequently, Optuna has been widely adopted to improve the performance of gradient boosting algorithms, including LightGBM, across various prediction tasks.

Even in light of these advances, a large proportion of the studies conducted thus far rely exclusively on standard performance indicators Accuracy, Precision, Recall, F1-score, and ROC-AUC to evaluate diabetes prediction models. Yet in healthcare settings, strong predictive accuracy on its own falls short, since clinicians must also grasp why a model arrives at a particular prediction before applying it within clinical decision-making. This gap has given rise to Explainable Artificial Intelligence (XAI) as a prominent area of research, oriented toward making such models more transparent. Of the many XAI techniques available, SHapley Additive exPlanations (SHAP) has gained particularly wide traction, largely because it is capable of delivering both global and local insights into how features contribute to predictions [8]. Prior research by Netayawijit et al. [9], Tran et al. [10], Islam et al. [11], and Pang [6] has shown that pairing SHAP with machine learning models not only boosts transparency but also helps to reveal which risk factors carry the greatest weight in relation to diabetes. Although explainability improves transparency by clarifying how prediction models reach their decisions, it does not guarantee that those decisions are equitable across different population groups. A model may provide interpretable predictions while still exhibiting systematic disparities among demographic groups. Therefore, explainability should be complemented by fairness evaluation to support the development of trustworthy artificial intelligence in healthcare.

Another growing concern in healthcare artificial intelligence is algorithmic fairness. Machine learning models may exhibit different predictive performance across demographic groups such as sex, age, and race, potentially introducing bias into healthcare decision-making. Liu *et al.* [12], Dang *et al.* [13], and García-Galindo *et al.* [14] emphasized that fairness evaluation has become a fundamental component of *Trustworthy Artificial Intelligence*. Nevertheless, fairness assessment remains considerably less explored in diabetes prediction studies than predictive performance and explainability.

Based on the existing literature, several research gaps can be identified. First, most diabetes prediction studies continue to emphasize predictive performance while evaluating models only through conventional classification metrics [3], [5], [6]. Second, studies incorporating SHAP generally focus on identifying influential predictors without assessing whether model performance remains consistent across different demographic groups [9], [10], [11]. Third, although population-scale datasets such as the Behavioral Risk Factor Surveillance System (BRFSS) have become increasingly important for diabetes prediction because they capture diverse demographic, behavioral, and clinical characteristics [5], [15], most previous studies relied on earlier BRFSS releases or other public datasets. The BRFSS 2024 dataset provides the most recent nationwide health survey data, reflecting current population characteristics and risk factor distributions. However, studies that comprehensively evaluate predictive performance, explainability, and fairness using this latest dataset remain scarce. These limitations indicate the need for a more comprehensive evaluation framework capable of producing prediction models that achieve high accuracy while also being interpretable and transparent.

In response to the gaps highlighted in prior literature, the present research puts forward a diabetes risk prediction model that draws on the BRFSS 2024 dataset and relies on the LightGBM algorithm, further refined through Optuna optimization. Departing from earlier studies, the proposed model is examined across three complementary dimensions: predictive performance, assessed via Accuracy, Balanced Accuracy, Precision, Recall, F1-score, and ROC-AUC; explainability, examined through SHAP; and fairness, evaluated using Fairlearn with respect to sensitive demographic attributes, specifically sex and race. By bringing these three evaluative dimensions together, this research contributes a holistic framework for building trustworthy machine learning models geared toward diabetes risk prediction.

This research is primarily distinguished by its implementation of a Trustworthy Artificial Intelligence evaluation framework for diabetes risk prediction, utilizing the large-scale BRFSS 2024 dataset. The proposed framework integrates predictive performance, explainability, and fairness within a single evaluation pipeline, thereby yielding models that achieve strong accuracy alongside interpretability and fairness. Furthermore, this research contributes to Sustainable Development Goal (SDG) 3: Good Health and Well-being by supporting early diabetes detection and SDG 10: Reduced Inequalities through fairness evaluation aimed at minimizing demographic bias in predictive models. The findings are expected to provide practical insights for developing reliable machine learning-based decision support systems in healthcare.

2. RESEARCH METHODOLOGY

2.1 Research Object

The research object of this study is the prediction of diabetes risk using the Behavioral Risk Factor Surveillance System (BRFSS) 2024 dataset published by the Centers for Disease Control and Prevention (CDC). BRFSS is an annual nationwide health survey conducted across the United States to collect information regarding health conditions, health-related behaviors, chronic disease risk factors, and access to healthcare services. The BRFSS 2024 dataset was selected because it represents the most recent nationally representative survey available, containing 454,774 respondents with comprehensive demographic, socioeconomic, lifestyle, and clinical information relevant to diabetes risk prediction. Compared with earlier BRFSS releases, the 2024 dataset provides the latest population characteristics, making it more suitable for developing predictive models intended to support current public health surveillance and

decision-making. The target variable in this study is diabetes status, which is categorized into three classes: non-diabetes, pre-diabetes, and diabetes.

2.2 Research Workflow

This study was conducted through a set of systematic phases intended to build a diabetes risk prediction model marked by strong predictive performance, interpretability, and fairness. The workflow applied throughout this research adheres to the typical stages found in machine learning development, comprising data collection, data preprocessing, model development, hyperparameter optimization, model evaluation, and the interpretation of results [16]. This process starts with gathering the 2024 Behavioral Risk Factor Surveillance System (BRFSS) dataset, proceeds through data preprocessing and the development of baseline models, continues with hyperparameter optimization carried out via Optuna, and is followed by predictive performance evaluation, model interpretation using SHapley Additive exPlanations (SHAP), and finally fairness assessment through Fairlearn. Figure 1 illustrates this workflow in its entirety.

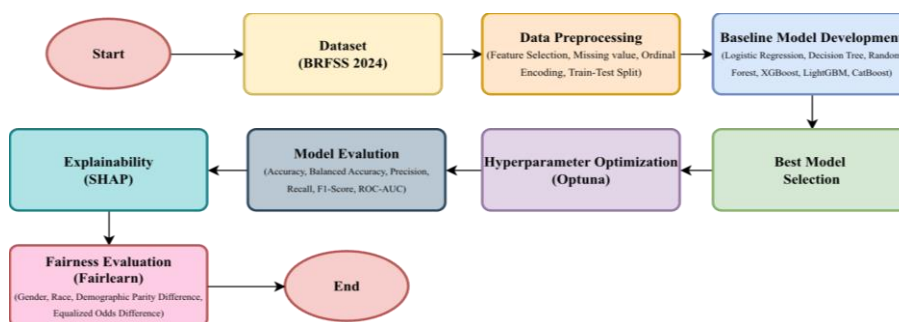


Figure 1. Research Workflow

Each stage serves a specific objective in the model development process. The preprocessing stage is intended to enhance data quality prior to model training by preparing the dataset for use with machine learning algorithms. Following this, a number of classification algorithms are constructed as baseline models, with the top-performing model subsequently refined through Optuna-based optimization. The optimized model is then evaluated using multiple classification metrics, interpreted through SHAP to determine how individual features influence the resulting predictions, and finally assessed using Fairlearn to examine potential prediction bias across sensitive attributes, namely gender and race.

2.3 Dataset

The dataset employed in this research is the Behavioral Risk Factor Surveillance System (BRFSS) 2024 dataset, released by the Centers for Disease Control and Prevention (CDC). BRFSS is a yearly, population-based health survey administered throughout the United States to gather data on health status, health-related behaviors, risk factors for chronic diseases, and access to healthcare services. The dataset was selected because it covers a large population and provides comprehensive information on demographic characteristics, health conditions, lifestyle behaviors, and healthcare access that are relevant to diabetes risk prediction [1], [15].

The BRFSS 2024 dataset contains 454,774 respondent records with 31 variables. After the feature selection process, 21 predictor variables and one target variable were retained for model development. The target variable represents diabetes status, which is classified into three categories: Non-diabetic, Pre-diabetic, and Diabetic. The predictor variables were grouped into several categories according to their characteristics, as presented in Table 1.

Table 1. Categories of Variables Used in This Study

Category	Variable
Demographic	Age, Sex, Race, Marital Status, Education, Income
Health Condition	Body Mass Index (BMI), General Health, Physical Health Days, Mental Health Days, Poor Health Days
Lifestyle	Exercise, Smoked 100 Cigarettes, Drinks Alcohol
Medical Record	Stroke, Heart Attack, Coronary Heart Disease
Healthcare Access	Cost Barrier, Last Checkup, Personal Doctor
Target Variable	Diabetes Status (Non-diabetic, Pre-diabetic, Diabetic)

2.4 Data Preprocessing

Data preprocessing was carried out with the aim of enhancing data quality prior to building the machine learning model. This stage aims to reduce data inconsistencies, handle missing values, and transform the dataset into a format compatible with the classification algorithms [2], [17]. The preprocessing pipeline consists of four main stages: feature selection, missing value handling, data transformation, and data splitting, as described in the following subsections.

2.4.1 Feature Selection

To eliminate variables that were irrelevant, redundant, or made no meaningful contribution to the predictive modeling process, a feature selection procedure was applied. Of the initial 31 variables in the BRFSS 2024 dataset, nine variables were excluded, resulting in 21 predictor variables used for model development. The retained variables represent demographic characteristics, health conditions, lifestyle factors, medical history, and healthcare access, all of which are considered relevant for diabetes risk prediction.

2.4.2 Missing Value Handling

After feature selection, several variables still contained missing values. Instead of removing incomplete records through listwise deletion, this study employed data imputation to preserve the large number of available observations and minimize unnecessary information loss. Considering the large sample size of the BRFSS 2024 dataset and the relatively low proportion of missing values in the selected variables, simple imputation techniques were considered appropriate for maintaining data completeness while avoiding substantial distortion of the original data distribution. Missing values in numerical variables were imputed using the median because this measure is less sensitive to extreme values and provides a robust estimate of central tendency. For categorical variables, missing values were replaced with the mode to preserve the most frequently observed category. Similar imputation strategies have been widely adopted in healthcare machine learning studies due to their simplicity, computational efficiency, and ability to retain the full dataset for model development [2], [17]. Following the imputation process, no missing values remained in the dataset, allowing all observations to be included in subsequent analyses.

2.4.3 Data Transformation

The majority of machine learning algorithms are unable to handle categorical variables in their raw form. Therefore, Ordinal Encoding was applied to transform categorical variables into numerical representations. Data transformation was performed after the missing value handling process to ensure that all variables were available in a numerical format suitable for model training.

2.4.4 Data Splitting

As the final stage of preprocessing, the dataset was divided into training and testing sets using an 80:20 ratio, with a stratified train-test split approach applied to maintain the target variable's class distribution in each subset. Model development and hyperparameter optimization were carried out using the training set, whereas the testing set was reserved for evaluating how well the trained model performed on unseen data.

2.4.5 Handling Class Imbalance

The BRFSS 2024 dataset exhibits a highly imbalanced class distribution, where the majority of respondents belong to the non-diabetes class, while the pre-diabetes class represents only a small proportion of the dataset. Such imbalance may bias machine learning models toward the majority class, resulting in poor predictive performance for minority classes. Rather than applying oversampling or undersampling techniques, this study addressed class imbalance by incorporating class weighting during model training. Logistic Regression, Decision Tree, Random Forest, and LightGBM were trained using the `class_weight="balanced"` parameter, whereas CatBoost employed the `auto_class_weights="Balanced"` option. These weighting strategies automatically assign larger penalties to minority-class misclassifications, enabling the learning algorithm to pay greater attention to underrepresented classes while preserving the original data distribution. In addition, macro F1-score was adopted as the optimization objective during hyperparameter tuning because it gives equal importance to each class regardless of its frequency, making it more suitable than overall accuracy for imbalanced multiclass classification problems.

2.5 Classification Model Development

In this study, a number of classification models were constructed to function as baseline models, with the aim of pinpointing which algorithm delivers the most robust predictive performance for diabetes risk prediction. The algorithms under consideration include Logistic Regression, Decision Tree, Random Forest, Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and CatBoost. Such a selection was intended to capture a broad cross-section of machine learning approaches, ranging from linear models (Logistic Regression) and decision tree-based models (Decision Tree) to bagging ensembles (Random Forest) and gradient boosting techniques (XGBoost, LightGBM, and CatBoost), each of which has seen considerable use in earlier research on diabetes prediction [3], [4], [5], [6], [18].

Baseline performance was determined by training each model on the preprocessed training dataset while retaining its default hyperparameter settings. This baseline evaluation offers a consistent point of reference for comparing the various algorithms prior to hyperparameter optimization. In line with this approach, earlier studies on healthcare classification have likewise assessed machine learning models across multiple performance metrics to guarantee a thorough evaluation [19]. Model performance in this study was gauged based on Accuracy, Balanced Accuracy, Precision, Recall, F1-score, and ROC-AUC, after which the algorithm with the strongest results was chosen to undergo hyperparameter optimization using Optuna.

2.6 Hyperparameter Optimization Using Optuna

Once all baseline models had been evaluated, the algorithm demonstrating the strongest performance was subsequently optimized through Optuna in order to determine the most effective combination of hyperparameters. Optuna functions as a Bayesian optimization framework that employs the Tree-structured Parzen Estimator (TPE) algorithm to intelligently explore the hyperparameter space and identify promising parameter combinations [7]. In comparison to conventional optimization approaches such as Grid Search and Random Search, Optuna offers a more efficient and adaptive search mechanism, especially when dealing with high-dimensional hyperparameter spaces.

For this study, the hyperparameter optimization process was applied specifically to the selected LightGBM model. The hyperparameters subjected to tuning consisted of `learning_rate`, `n_estimators`, `max_depth`, `num_leaves`, `min_child_samples`, `subsample`, `colsample_bytree`, `reg_alpha`, and `reg_lambda`. The optimization was directed toward maximizing the macro F1-score, given that this metric treats all classes with equal weight irrespective of their distribution. Considering the class imbalance present in the BRFSS 2024 dataset, the macro F1-score was regarded as a more suitable metric than overall accuracy for steering the hyperparameter optimization process and enhancing classification performance across all classes. The search ranges of the optimized hyperparameters are presented in Table 2.

Table 2. Hyperparameter Search Space for Optuna Optimization

Hyperparameter	Search Range
<code>learning_rate</code>	0,01 - 0,20
<code>n_estimators</code>	200 - 700
<code>max_depth</code>	4 - 12
<code>num_leaves</code>	31 - 200
<code>min_child_samples</code>	20 - 100
<code>subsample</code>	0,70 - 1,00
<code>colsample_bytree</code>	0,70 - 1,00
<code>reg_alpha</code>	0 - 3
<code>reg_lambda</code>	0 - 5

2.7 Model Performance Evaluation

Each classification model's performance was assessed using the testing dataset. Given that the BRFSS 2024 dataset displays an imbalanced class distribution, depending solely on overall accuracy could yield a misleading picture of model performance. For this reason, several evaluation metrics were applied to achieve a more thorough assessment, namely Accuracy, Balanced Accuracy, Precision, Recall, F1-score, and the Receiver Operating Characteristic–Area Under the Curve (ROC-AUC). Accuracy reflects the overall proportion of instances correctly classified, while Balanced Accuracy calculates the average recall across all classes, rendering it more appropriate for datasets with class imbalance. The macro-average approach was applied in computing Precision, Recall, and F1-score, as this approach gives equal weight to every class irrespective of how frequently it occurs. Such an approach guarantees that performance on minority classes receives a fair evaluation. ROC-AUC was further utilized to measure how well the model differentiated among the three diabetes classes. The evaluation metrics were calculated using Equations (1)–(5).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Balanced\ Accuracy = \frac{1}{C} \sum_{i=1}^C Recall_i \quad (2)$$

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (5)$$

2.8 Explainability Using SHAP

In an effort to make the model more interpretable, this study incorporates SHapley Additive exPlanations (SHAP) as its chosen Explainable Artificial Intelligence (XAI) technique. Rooted in Shapley values from cooperative game theory, SHAP determines how substantially each feature shapes the model's predictions. Under this approach, every feature is assigned a contribution score that shows the degree to which it either raises or lowers the predicted likelihood of diabetes. The decision to use SHAP was driven by its ability to furnish both global and local explanations, which in turn allows for a clearer view of the reasoning behind the model's outputs [8], [20], [21].

Interpretation of the model relied on TreeExplainer, a tool built specifically for tree-based machine learning models. To develop a well-rounded picture of the model's behavior, four types of SHAP visualizations were produced: the SHAP Summary Plot, SHAP Feature Importance, SHAP Dependence Plot, and SHAP Force Plot. Collectively, these visualizations clarify both the overall significance of predictor variables and the way individual features shape specific predictions, ultimately strengthening the transparency and interpretability of the diabetes risk prediction model proposed in this study.

2.9 Fairness Evaluation Using Fairlearn

This study extends beyond predictive performance by also evaluating the fairness of the proposed model, examining whether its performance holds consistently across various population groups. Fairness evaluation was conducted using the Fairlearn library by considering two sensitive attributes: gender and race. This evaluation aims to identify potential algorithmic bias toward specific demographic groups, thereby ensuring that the developed model produces more equitable and trustworthy predictions [12], [13], [14].

Fairness assessment in this study relied on two metrics, namely Demographic Parity Difference (DPD) and Equalized Odds Difference (EOD). The former, DPD, quantifies the disparity between sensitive groups in terms of how likely each is to receive a positive prediction, whereas the latter, EOD, captures the variation in true positive rate (TPR) and false positive rate (FPR) observed across these groups. As the values of both metrics decrease, this points toward diminished prediction bias and, by extension, a model that can be considered fairer.

3. RESULT AND DISCUSSION

3.1 Data Preprocessing Result

Data preprocessing was performed to prepare the BRFS 2024 dataset for classification model development. The preprocessing pipeline included feature selection, missing value imputation, categorical data encoding, and dataset splitting, resulting in a clean and structured dataset containing 21 predictor variables for subsequent model training.

Missing data were resolved through the application of median imputation for numerical features and mode imputation for categorical features, a process that allowed every observation to be kept intact. The dataset, once processed, was then split into 363,819 training samples and 90,955 testing samples by means of an 80:20 stratified train-test split, which preserved a consistent class distribution of the target variable across both subsets. This finalized dataset subsequently formed the basis for baseline model evaluation and hyperparameter optimization.

3.2 Baseline Model Evaluation

As an initial step, baseline models were established using six classification algorithms: Decision Tree, Random Forest, Logistic Regression, XGBoost, LightGBM, and CatBoost. All models underwent training on the same training dataset and were then evaluated against the testing dataset by means of Accuracy, Balanced Accuracy, Precision, Recall, F1-score, and ROC-AUC. Employing this range of metrics helped ensure a thorough evaluation of model performance, particularly given the imbalanced class distribution characterizing the BRFS 2024 dataset.

Table 3. Performance Comparison of Baseline Models

Model	Accuracy	Balanced Accuracy	Precision	Recall	F1 Score	ROC-AUC
Decision Tree	0.745479	0.399567	0.394380	0.399567	0.396538	0.563464
XGBoost	0.834490	0.376782	0.456771	0.376782	0.381387	0.774385
CatBoost	0.835138	0.376645	0.459968	0.376645	0.381246	0.777269
LightGBM	0.835710	0.375514	0.463103	0.375514	0.379660	0.778980
Random Forest	0.828025	0.373410	0.439778	0.373410	0.376742	0.737807
Logistic Regression	0.832027	0.372466	0.447099	0.372466	0.375143	0.767692

Table 3 shows that LightGBM delivered the strongest overall performance, recording the highest Accuracy, Precision, and ROC-AUC among all the models evaluated, with CatBoost and XGBoost trailing closely behind. Decision Tree, on the other hand, consistently yielded the weakest results across the majority of evaluation metrics. These findings indicate that gradient boosting algorithms are more effective than single-tree and linear models in capturing the complex nonlinear relationships among demographic, lifestyle, and clinical variables within the BRFS 2024 dataset.

Although the boosting-based models attained high Accuracy scores, all baseline models exhibited comparatively low values for Balanced Accuracy, Recall, and F1-score. This finding indicates that the models remained biased toward the majority class and struggled to correctly identify minority classes, especially the Pre-diabetic category. Consequently, overall Accuracy alone was insufficient to reflect the true predictive capability of the models on the imbalanced dataset. This observation highlights the importance of incorporating class-balanced evaluation metrics, such as Balanced Accuracy and macro F1-score, when assessing multiclass medical classification models. Taken together, LightGBM exhibited the most competitive results across the evaluation metrics, leading to its selection for the subsequent hyperparameter optimization stage using Optuna. Besides its predictive performance,

LightGBM has demonstrated high computational efficiency and scalability for large-scale tabular learning tasks, making it well suited for handling high-dimensional healthcare datasets such as BRFSS 2024 [22].

3.3 Hyperparameter Optimization Using Optuna

Even though LightGBM delivered the strongest results during the baseline evaluation, the model continued to show limited capability in classifying minority classes, as reflected by its relatively low Balanced Accuracy and macro F1-score. Therefore, hyperparameter optimization was performed using Optuna, with the aim of enhancing the model's generalization ability, particularly for the imbalanced class distribution of the BRFSS 2024 dataset.

The optimization process was conducted on nine key hyperparameters of the LightGBM model using Optuna with 30 optimization trials. The optimization objective was to maximize the macro F1-score, which assigns equal importance to each class regardless of class distribution. This metric was selected because the BRFSS 2024 dataset exhibits class imbalance, making the macro F1-score more suitable than overall accuracy for guiding the optimization process. The best hyperparameter configuration obtained from the optimization is presented in Table 4.

Table 4. Best Hyperparameter Configuration Obtained Using Optuna

Hyperparameter	Value
Learning Rate	0.1613
Number of Estimators	436
Maximum Depth	12
Number of Leaves	190
Minimum Child Samples	48
Subsample	0.9169
Colsample Bytree	0.9123
Reg Alpha	1.1739
Reg Lambda	1.1615

The optimized hyperparameter configuration was subsequently used to retrain the LightGBM model before evaluating its performance on the testing dataset. The comparison between the baseline and optimized models is presented in Table 5.

Table 5. Performance Comparison of LightGBM Before and After Hyperparameter Optimization

Model	Accuracy	Balanced Accuracy	Precision	Recall	F1 Score	ROC-AUC
LightGBM	0.835710	0.375514	0.463103	0.375514	0.379660	0.778980
LightGBM + Optuna	0.702336	0.489079	0.428301	0.489079	0.433297	0.729204

As shown in Table 5, hyperparameter optimization produced different effects across the evaluation metrics. Although overall Accuracy decreased from 83.57% to 70.23%, and ROC-AUC declined from 0.7790 to 0.7292, the optimized model achieved substantial improvements in metrics that are more suitable for imbalanced multiclass classification. Specifically, Balanced Accuracy increased from 0.3755 to 0.4891, while the macro F1-score improved from 0.3797 to 0.4333. These results indicate that the optimized model became more effective at recognizing all diabetes categories, particularly the underrepresented classes that were not well identified by the baseline model.

These findings illustrate the trade-off between maximizing overall Accuracy and improving balanced performance across all classes. Because the BRFSS 2024 dataset is highly imbalanced, optimizing the model based solely on Accuracy tends to favor the majority non-diabetes class while providing limited sensitivity to minority classes, particularly pre-diabetes. In contrast, optimizing the model using the macro F1-score encourages more balanced learning by assigning equal importance to every class, thereby improving the model's ability to detect underrepresented cases.

From a clinical perspective, this trade-off is acceptable because diabetes risk prediction is primarily intended as an early screening tool rather than a diagnostic replacement. Missing individuals who are at risk of diabetes or pre-diabetes (false negatives) may delay preventive intervention and increase the likelihood of disease progression. Therefore, improving Balanced Accuracy and macro F1-score is clinically more meaningful than maximizing overall Accuracy, as these metrics better reflect the model's capability to identify patients across all diabetes categories instead of predominantly recognizing the majority class.

These findings are consistent with previous studies by Islam et al. [11], which reported that hyperparameter optimization can improve minority-class recognition even when overall Accuracy does not increase. Similar observations have also been reported in healthcare classification studies, where Balanced Accuracy and macro F1-score are considered more informative than Accuracy for evaluating models trained on imbalanced clinical datasets. Based on these findings, the Optuna-optimized LightGBM model was selected as the final prediction model because it achieved a better balance between overall predictive performance and minority-class recognition. The optimized model was subsequently used for explainability analysis using SHAP and fairness evaluation using Fairlearn, allowing the proposed framework to be assessed not only in terms of predictive capability but also with respect to transparency and algorithmic fairness.

3.4 Model Explainability Using SHAP

Ensemble-based machine learning models, including LightGBM, typically deliver strong predictive performance yet frequently suffer from limited interpretability, given that the reasoning behind their decision-making process is not easily understood. To improve model transparency, this research applied SHapley Additive exPlanations (SHAP) to determine the extent of each feature's contribution to the model's predictions. The analysis was performed on the optimized LightGBM model using SHAP Summary Plot and SHAP Feature Importance.

3.4.1 SHAP Summary Plot

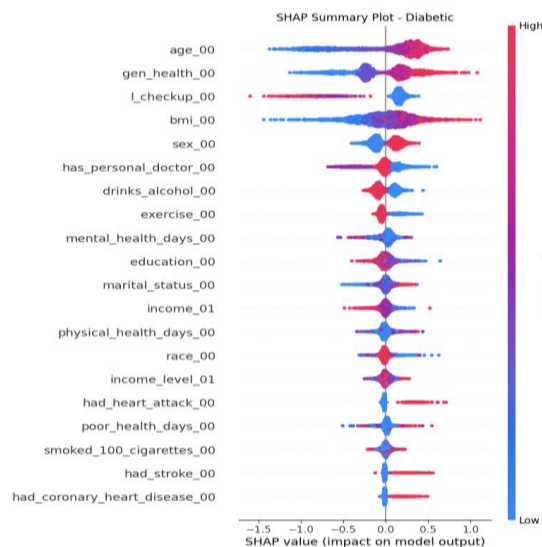


Figure 2. SHAP Summary Plot

Figure 2 presents the distribution of SHAP values for all predictor features in the testing dataset. Features positioned higher in the plot exert a greater overall influence on the model predictions than those located lower. Each point represents an individual observation, whereas its color corresponds to the original feature value, with warmer colors indicating relatively higher values and cooler colors indicating relatively lower values. Together, the horizontal distribution and color gradient illustrate both the magnitude and direction of each feature's contribution to the prediction outcome, providing a global interpretation of the optimized LightGBM model.

The SHAP Summary Plot indicates that Body Mass Index (BMI) is the most influential predictor, followed by Age, Physical Health Days, Mental Health Days, Income Level, and General Health. Higher BMI and older age generally produce positive SHAP values, indicating an increased probability of being classified into the diabetic category. This finding suggests that the optimized LightGBM model primarily relies on well-established metabolic risk factors when generating predictions. Moreover, the importance of self-reported physical and mental health indicators implies that the model captures broader aspects of an individual's health status rather than relying solely on physiological measurements. The contribution of socioeconomic variables, such as Income Level, further indicates that diabetes risk prediction is influenced by multiple interacting demographic, clinical, and social determinants.

3.4.2 SHAP Feature Importance

To gain further insight into the global contribution of each feature, a SHAP Feature Importance analysis was conducted. In contrast to the SHAP Summary Plot, which depicts the distribution of SHAP values across individual observations, SHAP Feature Importance ranks variables according to their mean absolute SHAP values, providing an overall measure of feature importance across the entire dataset.

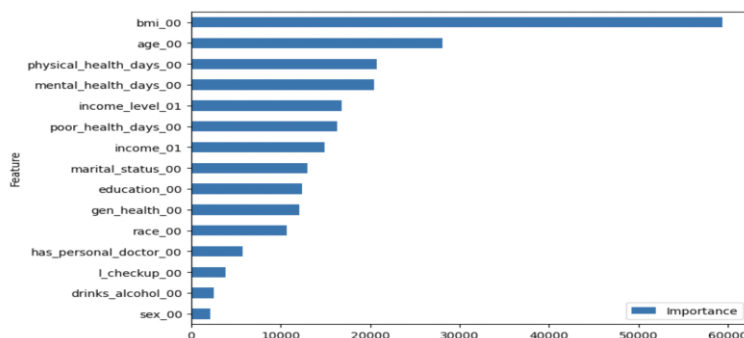


Figure 3. SHAP Feature Importance

As shown in Figure 3, only a limited number of features contribute substantially to the prediction process. The ten most influential features are summarized in Table 6.

Table 6. Top 10 Most Important Features Based on Mean Absolute SHAP Values

Rank	Feature	Mean Absolute SHAP Value
1	BMI	59.384
2	Age	28.144
3	Physical Health Days	20.708
4	Mental Health Days	20.454
5	Income Level	16.862
6	Poor Health Days	16.349
7	Income	14.899
8	Marital Status	12.998
9	Education	12.397
10	General Health	12.096

According to Table 6, BMI is the most influential feature, with a mean absolute SHAP value of 59.384, followed by Age (28.144), Physical Health Days (20.708), Mental Health Days (20.454), and Income Level (16.862). The substantial difference between BMI and the remaining variables indicates that body mass contributes disproportionately to the model's decision-making process. These findings suggest that the optimized LightGBM model primarily distinguishes diabetes risk based on metabolic characteristics while simultaneously incorporating information related to general health status and socioeconomic conditions. This finding also demonstrates the value of the BRFSS survey, as it captures not only clinical indicators but also behavioral, demographic, and socioeconomic information, enabling the model to identify diabetes risk from multiple complementary perspectives.

These findings are consistent with previous studies by Pang [6], Netayawijit et al. [9], and Islam et al. [11], which likewise identified BMI and age as the dominant predictors of diabetes. The agreement with previous studies indicates that the proposed model learns clinically meaningful patterns that are consistent with established epidemiological evidence rather than dataset-specific relationships. Furthermore, the importance assigned to Physical Health Days, Mental Health Days, and Income Level suggests that diabetes risk is not determined solely by physiological factors but is also associated with an individual's overall health condition and socioeconomic environment. Overall, the SHAP Feature Importance analysis demonstrates that the optimized LightGBM model generates predictions based on clinically relevant and epidemiologically plausible risk factors. This enhances the transparency of the prediction process by allowing healthcare practitioners to understand which variables drive the model's decisions, thereby increasing confidence in its potential application as a decision-support tool for early diabetes risk assessment.

3.5 Fairness Evaluation

In addition to predictive performance and interpretability, fairness represents a fundamental component of trustworthy machine learning, particularly in healthcare applications where prediction outcomes may influence clinical decisions. A highly accurate model is not necessarily suitable for practical deployment if it systematically performs better for certain demographic groups than for others. Therefore, the fairness of the optimized LightGBM model was evaluated using two sensitive demographic attributes, namely gender and race, to assess whether the proposed framework provides equitable prediction performance across different population groups.

The fairness evaluation was conducted using two complementary approaches. First, predictive performance was compared across demographic groups using Accuracy, Precision, Recall, and F1-score. Second, group fairness was evaluated by means of two metrics implemented within Fairlearn: Demographic Parity Difference (DPD) and Equalized Odds Difference (EOD). Lower values of both metrics indicate smaller disparities between groups and therefore a lower level of algorithmic bias.

3.5.1 Fairness Evaluation by Gender

The first fairness analysis was performed using gender as the sensitive attribute to determine whether the proposed model produced comparable predictive performance for male and female respondents. The results are presented in Table 7.

Table 7. Model Performance by Gender

Gender	Sample	Accuracy	Precision	Recall	F1-Score
Male	47.836	0,7049	0,4257	0,4849	0,4312
Female	43.119	0,6995	0,4311	0,4928	0,4350

As shown in Table 7, the predictive performance remains highly consistent across male and female respondents. The difference in Accuracy between the two groups is only approximately 0.5 percentage points, while

Precision, Recall, and F1-score also exhibit only marginal variation. These findings indicate that the optimized LightGBM model performs comparably regardless of gender, suggesting that the learning process is not disproportionately influenced by this demographic attribute. Such consistency is particularly important in healthcare applications, where unequal predictive performance between demographic groups could potentially contribute to disparities in clinical risk assessment.

To further assess fairness, Demographic Parity Difference and Equalized Odds Difference were calculated, as summarized in Table 8.

Table 8. Fairness Metrics by Gender

Metric	Value
Demographic Parity Difference	0,0379
Equalized Odds Difference	0,0379

The optimized model achieved a Demographic Parity Difference of 0.0379 and an Equalized Odds Difference of 0.0569, with both values remaining close to zero. These results indicate only minor differences in prediction outcomes and error rates between male and female respondents. From the perspective of group fairness, the model does not exhibit substantial demographic disparities, suggesting that it can provide relatively equitable diabetes risk predictions across gender groups. This characteristic enhances the trustworthiness of the proposed framework and supports its potential application in population-level diabetes screening.

3.5.2 Fairness Evaluation by Race

Fairness was further evaluated across racial groups because the BRFSS dataset contains respondents from diverse racial backgrounds with substantially different sample sizes. The predictive performance for each racial group is presented in Table 9.

Tabel 9. Predictive Performance by Race

Race	Sample	Accuracy	Precision	Recall	F1-Score
White	67.184	0,7037	0,4205	0,4820	0,4242
Black	6.975	0,6360	0,4388	0,4798	0,4308
Hispanic	9.669	0,7264	0,4494	0,5038	0,4595
Asian	1.313	0,6695	0,4341	0,4777	0,4313
American Indian or Alaskan Native	415	0,6361	0,4195	0,4525	0,4204
Native Hawaiian or Other Pacific Islander	2.565	0,7700	0,4571	0,5189	0,4733
Multiracial	2.084	0,7404	0,4490	0,5122	0,4619
Other	750	0,6440	0,3957	0,4372	0,3930

As shown in Table 9 and Figure 4, predictive performance varies across racial groups. The Native Hawaiian or Other Pacific Islander group achieved the highest Accuracy (0.7700), followed by the Multiracial (0.7404) and Hispanic (0.7264) groups. In contrast, the American Indian or Alaskan Native, Black, and Other groups exhibited comparatively lower Accuracy values.

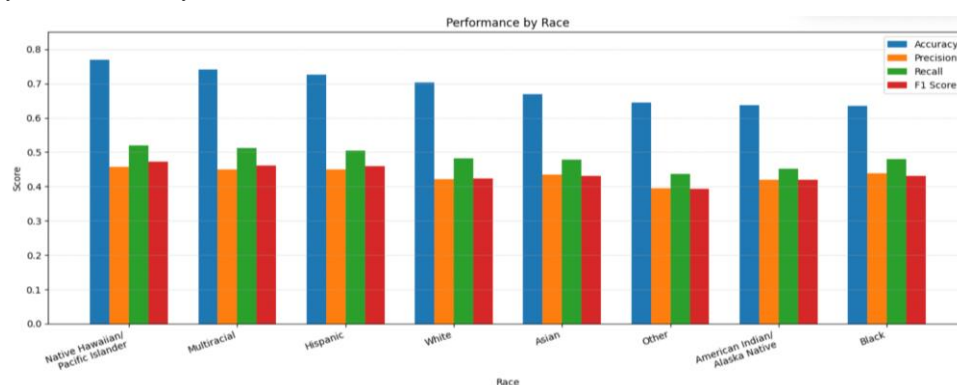


Figure 4. Predictive Performance Across Racial Groups

These differences are likely attributable to the unequal distribution of samples across racial groups in the BRFSS 2024 dataset rather than to systematic bias in the prediction model. The White population represents the largest proportion of respondents, whereas several minority groups contain substantially fewer observations. Consequently, the model has considerably fewer training examples from which to learn representative patterns for minority populations, making classification more challenging for these groups. This limitation is commonly encountered in large-scale healthcare datasets and should be considered when interpreting fairness across demographic subpopulations.

Nevertheless, the differences in Precision, Recall, and F1-score remain relatively modest across racial groups, indicating that the optimized model maintains reasonably stable predictive performance despite differences in sample availability. These findings suggest that the observed performance variation is more closely associated with data representation than with systematic algorithmic discrimination. Therefore, improving the representation of minority populations in future datasets may further reduce performance disparities while preserving the overall fairness of the prediction model.

3.5.3 Summary of Fairness Evaluation

Overall, the fairness evaluation demonstrates that the optimized LightGBM model achieves not only competitive predictive performance but also satisfactory fairness across the evaluated sensitive attributes. For gender, the low values of Demographic Parity Difference (0.0379) and Equalized Odds Difference (0.0569) indicate only minimal disparities between male and female respondents. Although predictive performance varies across racial groups, these differences are more likely attributable to the unequal representation of racial groups in the BRFSS 2024 dataset rather than to systematic algorithmic bias. Therefore, the proposed model is capable of providing relatively equitable predictions across diverse demographic populations while maintaining acceptable predictive performance.

The results further demonstrate that predictive performance, explainability, and fairness can be successfully integrated within a unified machine learning framework for diabetes risk prediction. By combining these complementary evaluation perspectives, the proposed framework produces a model that is not only accurate but also transparent and equitable, thereby supporting the principles of Trustworthy Artificial Intelligence. From a practical perspective, this integrated evaluation framework provides a stronger foundation for the development of clinical decision support systems and population-level diabetes screening tools while reducing the potential risk of demographic bias in machine learning-based healthcare applications. Nevertheless, the fairness evaluation in this study is limited to gender and race, and the unequal distribution of respondents across racial groups may influence the observed performance differences. Future studies should investigate additional sensitive attributes and explore bias mitigation techniques to further improve fairness in diabetes risk prediction models.

3.6 Discussion

The findings of this study demonstrate that Optuna-based hyperparameter optimization improved the ability of the LightGBM model to recognize minority diabetes classes. Compared with the baseline model, Balanced Accuracy increased from 0.3755 to 0.4891, while macro F1-score improved from 0.3797 to 0.4333, although overall Accuracy decreased from 0.8357 to 0.7023. This trade-off is expected in highly imbalanced multiclass datasets because the optimized model becomes less biased toward the majority non-diabetic class and achieves better discrimination of minority diabetes categories. Similar findings have been reported by Das et al. [3], Li et al. [4], and Rafie et al. [5], who demonstrated that ensemble learning methods provide strong predictive performance for diabetes prediction. However, unlike previous studies that primarily emphasized overall Accuracy, the present findings indicate that Balanced Accuracy and macro F1-score provide more appropriate evaluation metrics for imbalanced healthcare datasets, where correctly identifying pre-diabetic and diabetic individuals is more important than maximizing prediction accuracy for the majority class. Consequently, the proposed optimization strategy offers a more balanced evaluation of model performance and is better suited for supporting early diabetes risk identification in real-world clinical settings.

The explainability and fairness analyses further strengthen the trustworthiness of the proposed framework. Consistent with previous studies by Pang [6], Netayawijit et al. [9], and Islam et al. [11], SHAP analysis identified Body Mass Index (BMI), Age, Physical Health Days, Mental Health Days, and Income Level as the most influential predictors, indicating that the optimized LightGBM model learned clinically meaningful patterns consistent with established diabetes risk factors. Fairness evaluation also demonstrated highly consistent predictive performance across gender, with an Accuracy difference of only approximately 0.5%, while both Demographic Parity Difference and Equalized Odds Difference were 0.0379, indicating minimal demographic disparity. Although moderate performance variation was observed across racial groups, these differences are more likely attributable to unequal sample representation than to systematic algorithmic bias. Previous diabetes prediction studies have predominantly focused on improving predictive performance [3], [4], [5], whereas subsequent research incorporated explainability through SHAP and LIME to improve model transparency [6], [9], [11]. More recent studies have further highlighted fairness as an essential requirement for trustworthy healthcare artificial intelligence [12], [13], [14], [15]. Building upon these developments, the present study integrates predictive performance, explainability, and fairness within a single evaluation framework, thereby providing a more comprehensive, transparent, equitable, and clinically reliable approach for diabetes risk prediction.

4. CONCLUSION

This study developed a diabetes risk prediction model using the Behavioral Risk Factor Surveillance System (BRFSS) dataset through a machine learning approach. Among the evaluated baseline classifiers, LightGBM achieved the best overall performance and was subsequently optimized using Optuna. Although hyperparameter optimization reduced

overall Accuracy, it substantially improved Balanced Accuracy and macro F1-score, resulting in a more balanced classification performance for the imbalanced multiclass dataset. Model interpretability was enhanced through SHapley Additive exPlanations (SHAP), which identified Body Mass Index (BMI), Age, Physical Health Days, Mental Health Days, and Income Level as the most influential predictors of diabetes risk. Furthermore, fairness evaluation using Fairlearn demonstrated low demographic disparity across gender groups and relatively consistent predictive performance across different racial groups, indicating that the proposed model exhibits a low level of systematic bias. These findings demonstrate that improving predictive performance alone is insufficient for healthcare applications and that integrating explainability and fairness provides a more reliable basis for trustworthy clinical decision support. Consequently, the proposed evaluation framework offers a practical approach for developing machine learning-based diabetes risk screening systems that not only improve the identification of high-risk individuals but also support transparent and equitable decision-making in public health. Nevertheless, this study has several limitations, including the evaluation of fairness using only gender and race as sensitive attributes and the use of a single optimized classification algorithm. Future research should investigate additional sensitive attributes, such as education level and socioeconomic status, evaluate external datasets to assess model generalizability, and explore bias mitigation techniques as well as fairness-aware machine learning methods to further improve predictive performance, fairness, and clinical applicability.

5. DECLARATIONS

5.1 Author Contributions

Conceptualization: V.D.N., B.W.P., M.R.S., and D.R.I.;

Methodology: V.D.N., B.W.P., and M.R.S.;

Software: V.D.N.;

Validation: V.D.N., B.W.P., M.R.S., and D.R.I.;

Formal Analysis: V.D.N., B.W.P., and M.R.S.;

Investigation: V.D.N., B.W.P., and M.R.S.;

Resources: V.D.N.;

Data Curation: V.D.N.;

Writing Original Draft Preparation: V.D.N., B.W.P., and M.R.S.;

Writing Review and Editing: V.D.N., B.W.P., M.R.S., and D.R.I.;

Visualization: V.D.N.;

All authors have read and agreed to the published version of the manuscript.

5.2 Data Availability

The data presented in this study are available on request from the corresponding author.

5.3 Funding

This research and the publication of this manuscript were fully funded by the author's personal funds. No external funding was received for the research, preparation, or publication of this manuscript.

5.4 Institutional Review Board Statement

Not applicable.

5.5 Informed Consent Statement

Not applicable.

5.6 Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

5.7 Generative AI and AI-assisted Technologies in The Writing Process

The authors used Generative AI to improve the writing clarity of this paper. They reviewed and edited the AI-assisted content and took full responsibility for the final publication.

REFERENCES

- [1] D. Magliano and E. J. Boyko, *IDF Diabetes Atlas*, 11th ed. Brussels, Belgium: International Diabetes Federation, 2025.
- [2] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2017.

- [3] D. Das, Aayushman, S. Kumar, M. A. Hussain, and B. R. Reddy, “Diabetes prediction using ensemble learning techniques,” *Procedia Computer Science*, vol. 258, pp. 3155–3164, 2025, doi: 10.1016/j.procs.2025.04.573.
- [4] 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, 2024, doi: 10.1371/journal.pone.0311222.
- [5] 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, 2025, doi: 10.1186/s12889-025-24953-w.
- [6] K. Pang, “A comparative study of explainable machine learning models with Shapley values for diabetes prediction,” *Healthcare Analytics*, vol. 7, p. 100390, 2025, doi: 10.1016/j.health.2025.100390.
- [7] H. Almarzooq and U. bin Waheed, “Automating hyperparameter optimization in geophysics with Optuna: A comparative study,” *Geophysical Prospecting*, vol. 72, pp. 1778–1788, 2024, doi: 10.1111/1365-2478.13484.
- [8] A. M. Salih, Z. Raisi-Estabragh, I. B. Galazzo, P. Radeva, S. E. Petersen, K. Lekadir, and G. Menegaz, “A perspective on explainable artificial intelligence methods: SHAP and LIME,” *Advanced Intelligent Systems*, vol. 7, no. 1, Art. no. 2400304, 2025, doi: 10.1002/aisy.202400304.
- [9] 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, 2025, doi: 10.3390/healthcare13202588.
- [10] V. Q. Tran, Y. Choi, and H. Byeon, “Explainable stacking ensemble with feature tokenizer transformers for men's diabetes prediction,” *Journal of Men's Health*, vol. 20, no. 11, Art. no. 38, 2024, doi: 10.22514/jomh.2024.184.
- [11] Md. M. Islam, H. R. Rifat, Md. S. B. Shahid, A. Akhter, M. A. Uddin, and K. M. M. Uddin, “Explainable machine learning for efficient diabetes prediction using hyperparameter tuning, SHAP analysis, partial dependency, and LIME,” *Engineering Reports*, vol. 7, no. 1, p. e13080, 2025, doi: 10.1002/eng2.13080.
- [12] M. Liu *et al.*, “FAIM: Fairness-aware interpretable modeling for trustworthy machine learning in healthcare,” *Patterns*, vol. 5, no. 10, p. 101059, 2024, doi: 10.1016/j.patter.2024.101059.
- [13] V. N. Dang, V. M. Campello, J. Hernández-González, and K. Lekadir, “Empirical comparison of post-processing debiasing methods for machine learning classifiers in healthcare,” *Journal of Healthcare Informatics Research*, vol. 9, no. 3, pp. 465–493, 2025, doi: 10.1007/s41666-025-00196-7.
- [14] A. García-Galindo, M. López-De-Castro, and R. Armañanzas, “Fair prediction sets through multi-objective hyperparameter optimization,” *Machine Learning*, vol. 114, no. 1, Art. no. 27, 2025, doi: 10.1007/s10994-024-06721-w.
- [15] V. P. Nugraha and A. Febrian, “Comprehensive diabetes risk prediction using BRFSS data: Performance, explainability, fairness, and calibration,” *Journal of Applied Informatics and Computing*, vol. 10, no. 3, pp. 2357–2368, 2026, doi: 10.30871/jaic.v10i3.12740.
- [16] M. R. Sanjaya, A. Khoiriah, R. I. Heroza, and B. W. Putra, “Analysis of user reviews for the MyTelkomsel app using naïve Bayes and Random Forest methods,” *Jurnal Testing dan Implementasi Sistem Informasi*, vol. 4, no. 1, pp. 32–48, 2026, doi: 10.55583/jtisi.v4i1.2220.
- [17] M. H. H. Alkinani, A. M. Albugami, and M. R. Ibrahim, “Data preprocessing techniques in healthcare machine learning: A systematic review,” *Healthcare Analytics*, vol. 6, Art. no. 100336, 2024, doi: 10.1016/j.health.2024.100336.
- [18] G. R. Ashisha, X. A. Mary, E. G. M. Kanaga, J. Andrew, and R. J. Eunice, “Random oversampling-based diabetes classification via machine learning algorithms,” *International Journal of Computational Intelligence Systems*, vol. 17, no. 1, Art. no. 270, 2024, doi: 10.1007/s44196-024-00678-3.
- [19] B. W. Putra, “Deep neural network classification for ECG Arrhythmia Detection with Optimized Feature Autoencoder,” *Journal of Advanced Research in Applied Mechanics*, vol. 133, no. 1, pp. 78–86, 2025, doi: 10.37934/aram.133.1.7886.
- [20] M. Frasca, D. La Torre, G. Pravettoni, and I. Cutica, “Explainable and interpretable artificial intelligence in medicine: A systematic bibliometric review,” *Discover Artificial Intelligence*, vol. 4, no. 1, Art. no. 15, 2024, doi: 10.1007/s44163-024-00114-7.
- [21] J. Cateson, A. Lewin, and E. Williamson, “The application of explainable artificial intelligence (XAI) in electronic health record research: A scoping review,” *Digital Health*, vol. 10, Art. no. 20552076241272657, 2024, doi: 10.1177/20552076241272657.
- [22] C. Lokker, W. Abdelkader, E. Bagheri, et al., “Boosting efficiency in a clinical literature surveillance system with LightGBM,” *PLOS Digital Health*, vol. 3, no. 9, Art. no. e0000299, 2024, doi: 10.1371/journal.pdig.0000299.