Improving Diabetes Prediction Through IQR Outlier Treatment an Evaluation of XGBoost And LightGBM Models

Boda Sri Lakshmi Sravya*[1], G. Pranith[2]

[1]M. Tech Student, Department of CSE, Eluru College of Engineering and Technology, Duggirala (V), Pedavegi (M), Eluru- 534004,

[2]Assistant Professor, Department of CSE, Eluru College of Engineering and Technology, Duggirala (V), Pedavegi (M), Eluru- 534004.

DOI: https://doi.org/10.51583/IJLTEMAS.2026.150700113

Received: 01 August 2026; Accepted: 06 August 2026; Published: 20 August 2026

ABSTRACT

The global importance of Diabetes Mellitus (DM) is addressed by using advanced machine learning approaches to increase early detection and predictive precision. Due to the increasing prevalence of diabetes, especially in developing countries, the study shows that model applicability, algorithm choice, and long-term prediction accuracy are little understood. Dataset preprocessing—cleaning, scaling, and controlling outliers using the Interquartile Range (IQR) approach—is essential. Comparing XGBoost vs LightGBM shows differences in accuracy, precision, recall, F1 score, and AUC score. XGBoost with IQR preprocessing produce good accuracy, precision, and recall, making it a potential predictor. However, LightGBM has various performance indicators, highlighting the importance of considering environmental and application requirements when picking a model. To construct reliable diabetes prediction models, rigorous preprocessing and model validation are essential.

Keywords: Diabetes Mellitus, Type 2 Diabetes, Machine Learning, Deep Learning, Early Disease Prediction, Predictive Modeling, Clinical Decision Support System (CDSS).

INTRODUCTION

Diabetes mellitus is one of the most prevalent chronic metabolic disorders worldwide and represents a major public health concern. It is characterized by persistently elevated blood glucose levels resulting from inadequate insulin production, impaired insulin action, or a combination of both. Insulin, a hormone secreted by the pancreatic β-cells, plays a vital role in regulating blood glucose by facilitating its uptake into body tissues. When insulin production is insufficient or the body's cells become resistant to its action, glucose accumulates in the bloodstream, leading to hyperglycemia and various long-term health complications [1].

The global prevalence of diabetes has increased dramatically over the past few decades due to rapid urbanization, sedentary lifestyles, unhealthy dietary habits, obesity, and population aging. According to the World Health Organization (WHO), diabetes affects hundreds of millions of people worldwide and continues to be a leading cause of premature mortality and disability. Beyond its health consequences, diabetes imposes a considerable socioeconomic burden through increased healthcare expenditure, reduced workforce productivity, and diminished quality of life. Consequently, early diagnosis, continuous monitoring, and effective management are essential for minimizing disease progression and associated complications [2].

Diabetes is broadly classified into two major categories: Type 1 Diabetes Mellitus (T1DM) and Type 2 Diabetes Mellitus (T2DM). Type 1 diabetes is an autoimmune disease in which the immune system mistakenly destroys insulin-producing β-cells of the pancreas. This condition commonly develops during childhood or adolescence, although it may occur at any age. Individuals with Type 1 diabetes require lifelong insulin replacement therapy to maintain normal blood glucose levels.

In contrast, Type 2 diabetes is the most common form, accounting for approximately 90–95% of all diabetes

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

cases. It is primarily associated with insulin resistance, where body cells fail to respond effectively to insulin, combined with progressive pancreatic β-cell dysfunction. Several modifiable risk factors, including obesity, physical inactivity, poor dietary habits, smoking, and chronic stress, contribute significantly to the development of Type 2 diabetes. Genetic predisposition and family history also play important roles in increasing disease susceptibility. Appropriate lifestyle modifications, regular physical activity, healthy dietary practices, and pharmacological interventions can effectively delay or control the progression of Type 2 diabetes [3].

Given the increasing prevalence and associated health risks, diabetes has become a priority area for healthcare researchers and policymakers. Advances in medical technologies, artificial intelligence, machine learning, and predictive analytics have opened new opportunities for early diagnosis, personalized treatment planning, and continuous disease monitoring, ultimately improving patient outcomes and reducing healthcare costs.

Complications and Implications

Diabetes is associated with numerous acute and chronic complications that significantly affect an individual's health, quality of life, and life expectancy. Persistent hyperglycemia damages blood vessels, nerves, and vital organs, resulting in both microvascular and macrovascular complications. If left untreated or poorly controlled, diabetes can lead to severe and often irreversible health conditions.

Among the macrovascular complications, cardiovascular diseases remain the leading cause of mortality among diabetic patients. Individuals with diabetes are at a substantially higher risk of developing coronary artery disease, myocardial infarction, stroke, and peripheral arterial disease. These conditions arise primarily due to prolonged vascular damage caused by elevated blood glucose levels.

Microvascular complications include diabetic neuropathy, retinopathy, and nephropathy. Diabetic neuropathy results from nerve damage and may cause pain, tingling sensations, numbness, muscle weakness, and, in severe cases, foot ulcers that can necessitate limb amputation. Diabetic retinopathy affects the retinal blood vessels and is one of the leading causes of preventable blindness among adults. Similarly, diabetic nephropathy progressively impairs kidney function and may ultimately result in chronic kidney disease or end-stage renal failure requiring dialysis or kidney transplantation. These complications substantially increase healthcare costs and reduce patients' physical, emotional, and social well-being [4].

The prevention and effective management of diabetes require a comprehensive and multidisciplinary approach. Although Type 1 diabetes cannot currently be prevented, effective disease management can be achieved through lifelong insulin therapy, continuous glucose monitoring, regular medical consultations, balanced nutrition, and healthy lifestyle practices. Conversely, Type 2 diabetes can often be prevented or delayed through lifestyle modifications, including maintaining a healthy body weight, consuming a balanced diet, engaging in regular physical activity, avoiding tobacco use, and reducing excessive alcohol consumption.

For individuals already diagnosed with diabetes, treatment typically involves lifestyle interventions combined with oral antidiabetic medications or insulin therapy, depending on disease severity and patient-specific factors. Regular monitoring of blood glucose, glycated hemoglobin (HbA1c), blood pressure, lipid profile, kidney function, and eye health is essential for preventing disease progression and minimizing complications. Patient education, self-management strategies, and adherence to prescribed treatment plans play critical roles in achieving optimal glycemic control and improving long-term clinical outcomes. Public health initiatives, community awareness programs, and technological innovations such as wearable glucose monitoring systems, telemedicine platforms, and machine learning-based predictive models have significantly enhanced diabetes prevention, early detection, and personalized disease management. Continued research and the integration of advanced computational techniques into clinical practice are expected to further improve diagnostic accuracy, treatment effectiveness, and overall quality of life for individuals living with diabetes.

Realted Work

The increasing prevalence of diabetes has motivated researchers to develop intelligent computational models

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

for its early detection and diagnosis. Machine learning (ML) and deep learning (DL) techniques have demonstrated remarkable capabilities in analyzing complex healthcare datasets and assisting clinicians in identifying diabetic patients at an early stage. Numerous studies have utilized publicly available datasets, particularly the Pima Indians Diabetes Database (PIDD), to evaluate the effectiveness of various predictive models.

Saru et al. [5] proposed a machine learning-based diabetes prediction framework using the Pima Indians Diabetes Database. The study employed the WEKA data mining platform and incorporated a bootstrapping resampling technique to improve classification performance. Several ensemble learning algorithms were evaluated, achieving an overall prediction accuracy of 90.36%. The authors concluded that ensemble classifiers significantly improve prediction performance and suggested extending the framework to multiple disease prediction applications using larger healthcare datasets.

Sisodia et al. [6] investigated the effectiveness of three supervised machine learning algorithms, namely Decision Tree (DT), Support Vector Machine (SVM), and Naïve Bayes (NB), for diabetes diagnosis. Using the PIDD dataset obtained from the UCI Machine Learning Repository, the models were evaluated based on accuracy, precision, recall, and F1-score. Among the evaluated classifiers, the Naïve Bayes algorithm achieved the highest prediction accuracy of 76.30%. The study demonstrated the feasibility of machine learning techniques in supporting clinical decision-making and emphasized their applicability for diagnosing various chronic diseases.

Ayon et al. [7] presented a Deep Neural Network (DNN) model for diabetes prediction using the Pima Indians Diabetes dataset. The proposed architecture was trained using both five-fold and ten-fold cross-validation techniques to ensure model reliability. The DNN achieved an impressive prediction accuracy of 98.35% under five-fold cross-validation and 97.11% under ten-fold cross-validation, with corresponding high F1-scores and Matthews Correlation Coefficient (MCC) values. The study highlighted the superiority of deep learning models over conventional machine learning approaches in detecting diabetes at an early stage.

Naz et al. [8] conducted a comparative analysis of traditional machine learning algorithms and deep learning techniques for early diabetes prediction. Their findings revealed that deep learning models consistently outperformed conventional machine learning classifiers by achieving an overall prediction accuracy of 98%. The authors emphasized that accurate early diagnosis can significantly improve patient management and reduce the long-term complications associated with diabetes.

Sivaranjani et al. [9] explored diabetes prediction using Random Forest (RF) and Support Vector Machine (SVM) classifiers. The study applied Principal Component Analysis (PCA) for dimensionality reduction and feature optimization before classification. Experimental results indicated that the Random Forest classifier achieved an accuracy of 83%, whereas the Support Vector Machine obtained an accuracy of 81.4%. The study demonstrated that effective feature selection and dimensionality reduction contribute substantially to improving predictive performance, particularly for high-dimensional healthcare datasets.

Ashiquzzaman et al. [10] proposed a deep learning framework incorporating dropout regularization to overcome the problem of overfitting in diabetes prediction. The proposed neural network architecture was evaluated using the Pima Indians Diabetes dataset and achieved a prediction accuracy of 88.41%. The inclusion of dropout layers enhanced model generalization and produced more reliable prediction results compared with conventional neural network architectures.

Recent research has also focused on integrating wearable healthcare devices, Internet of Things (IoT) technologies, electronic health records (EHRs), and artificial intelligence techniques for continuous diabetes monitoring and prediction. These approaches enable real-time data collection, personalized healthcare recommendations, and early intervention, thereby improving patient outcomes. Furthermore, advances in explainable artificial intelligence (XAI), federated learning, and hybrid deep learning models have enhanced the interpretability, privacy, and scalability of diabetes prediction systems.

Overall, the literature demonstrates that machine learning and deep learning algorithms have significantly improved diabetes prediction accuracy, with reported accuracies ranging from 76.30% to 98.35%. However, challenges remain regarding dataset diversity, model interpretability, feature selection, class imbalance, and real-world clinical deployment. These limitations provide opportunities for developing more robust, explainable, and generalizable diabetes prediction frameworks.

Motivation

Diabetes continues to be one of the leading causes of morbidity and mortality worldwide, resulting in severe complications such as cardiovascular diseases, diabetic neuropathy, retinopathy, nephropathy, and lower-limb amputations. The growing incidence of diabetes, coupled with increasing healthcare costs, highlights the urgent need for accurate and efficient diagnostic systems.

Traditional diagnostic procedures often depend on laboratory investigations and expert clinical interpretation, which may delay early diagnosis. Recent advances in artificial intelligence, particularly machine learning and deep learning, provide promising opportunities for developing automated decision-support systems capable of predicting diabetes with high accuracy. Early identification of high-risk individuals enables timely medical intervention, lifestyle modification, and personalized treatment, ultimately reducing disease progression and improving patient quality of life. These factors collectively motivate the development of intelligent diabetes prediction models.

Research Gap

Although significant progress has been achieved in machine learning-based diabetes prediction, several research challenges remain unresolved.

- Most existing studies rely heavily on the Pima Indians Diabetes Database, limiting the generalizability of developed models to diverse populations.
- Existing prediction models often exhibit reduced performance when applied to heterogeneous clinical datasets collected from different geographical regions and healthcare systems.
- Many studies primarily focus on maximizing prediction accuracy while neglecting important aspects such as model interpretability, computational efficiency, robustness, and clinical applicability.
- Class imbalance, missing values, noisy medical records, and high-dimensional feature spaces continue to affect prediction reliability.
- Limited research has investigated hybrid optimization algorithms, explainable artificial intelligence techniques, and real-time healthcare data integration for diabetes prediction.

Addressing these research gaps will facilitate the development of more accurate, robust, explainable, and clinically applicable diabetes prediction systems suitable for real-world healthcare environments.

Objectives

The major objectives of this research are:

1. To conduct a comprehensive review of existing machine learning and deep learning techniques used for diabetes diagnosis and prediction.
2. To develop an optimized diabetes prediction framework that improves prediction accuracy, robustness, and computational efficiency through advanced feature selection and optimization techniques.
3. To evaluate and compare the proposed model with existing machine learning and deep learning algorithms using standard performance metrics such as accuracy, precision, recall, F1-score, specificity, and AUC.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

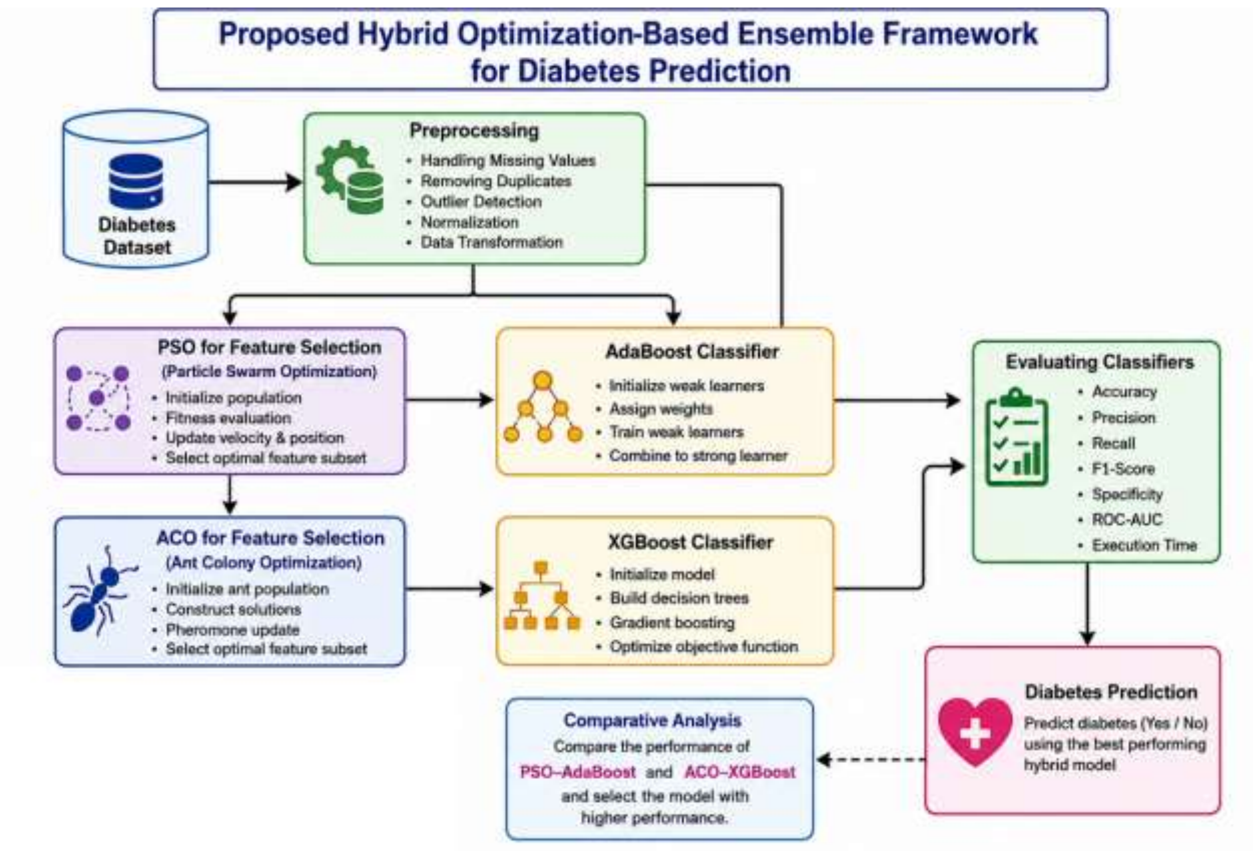
4. To identify the most significant clinical features contributing to diabetes prediction for improving model interpretability and supporting clinical decision-making.

5. To demonstrate the effectiveness of the proposed framework on benchmark diabetes datasets and establish its suitability for real-world healthcare applications.

RESEARCH METHODOLOGY

This research presents a machine learning-based framework for the early prediction of diabetes using advanced ensemble learning and optimization techniques. The proposed methodology begins with comprehensive preprocessing of the healthcare dataset to improve data quality and ensure reliable model performance. Data preprocessing involves handling missing values, removing duplicate records, detecting and treating outliers, normalizing feature values, and transforming the dataset into a suitable format for machine learning algorithms. These steps improve data consistency and enhance the predictive capability of the classification models. Following preprocessing, the dataset is divided into training and testing subsets for model development and evaluation. Feature selection and optimization techniques are then employed to identify the most informative attributes while reducing redundant and irrelevant features. This process minimizes computational complexity and improves classification accuracy.

The study investigates two hybrid ensemble learning models for diabetes prediction. The first model integrates Particle Swarm Optimization (PSO) with the AdaBoost classifier, where PSO is used to optimize model parameters and improve the classification performance of AdaBoost. The second model combines Ant Colony Optimization (ACO) **with** Extreme Gradient Boosting (XGBoost)**,** utilizing ACO to optimize feature selection and hyperparameters for enhancing the predictive capability of the XGBoost classifier.The performance of both hybrid models is evaluated using standard classification metrics, including accuracy, precision, recall, F1-score, specificity, Area Under the Receiver Operating Characteristic Curve (AUC-ROC), and computational time**.** Comparative analysis is conducted to determine the effectiveness of the proposed optimization-based ensemble approaches in accurately predicting diabetes. Figure 1 illustrates the proposed hybrid optimization-based ensemble framework for diabetes prediction. The framework integrates PSO–AdaBoost and ACO–XGBoost models to perform optimized feature selection, classification, and comparative performance evaluation for accurate diabetes prediction.



ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

Fig1: Proposed Hybrid Optimization-Based Ensemble Framework for Diabetes Prediction Using PSO–AdaBoost and ACO–XGBoost.

Dataset

The diabetes dataset used in this research was obtained from the Kaggle repository and is publicly available for machine learning applications. The dataset contains clinical and physiological measurements that are commonly used for diabetes prediction. It consists of 768 patient records with 8 input features and 1 target variable (Outcome), where the outcome indicates whether a patient has diabetes (1) or does not have diabetes (0).

The dataset includes several important attributes associated with diabetes diagnosis. These features are Pregnancies, representing the number of times a woman has been pregnant; Glucose, indicating plasma glucose concentration; Blood Pressure, representing diastolic blood pressure (mm Hg); Skin Thickness, measuring triceps skinfold thickness (mm); Insulin, representing 2-hour serum insulin (mu U/ml); Body Mass Index (BMI), calculated from an individual's weight and height; Diabetes Pedigree Function (DPF), estimating the hereditary risk of diabetes based on family history; and Age, representing the patient's age in years. The target variable, Outcome, identifies the presence or absence of diabetes.

The dataset is widely used as a benchmark for evaluating machine learning and deep learning algorithms due to its balanced clinical features and standardized structure. Before model development, the dataset undergoes preprocessing steps, including missing value treatment, normalization, and feature optimization, to improve prediction performance and model reliability.

Dataset Source: Kaggle Dataset: *Diabetes Dataset* https://www.kaggle.com/datasets/saurabh0007/diabetescsv

Table 1: Description of Dataset Attributes

Attribute	**Description**
Pregnancies	Number of pregnancies
Glucose	Plasma glucose concentration
Blood Pressure	Diastolic blood pressure (mm Hg)
Skin Thickness	Triceps skinfold thickness (mm)
Insulin	2-hour serum insulin (mu U/ml)
BMI	Body Mass Index (kg/m²)
DiabetesPedigreeFunction	Diabetes hereditary risk score
Age	Age of the patient (years)
Outcome	Diabetes status (0 = non-diabetic, 1 = Diabetic)

Machine Learning Models

IQR with XGBoost

The proposed IQR with XGBoost model integrates statistical outlier detection with the powerful Extreme Gradient Boosting (XGBoost) algorithm to improve diabetes prediction accuracy. Initially, the Interquartile

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

Range (IQR) method is employed during the preprocessing stage to detect and eliminate outliers from the healthcare dataset. Removing abnormal observations minimizes noise, enhances data quality, and improves the generalization capability of the classifier.

The Interquartile Range is calculated as:

$$IQR = Q_3 - Q_1 \qquad (Eq - 1)$$

where

- Q_1= First Quartile (25th percentile)
- Q_3= Third Quartile (75th percentile)

The lower and upper limits for outlier detection are computed as

$$LB = Q_1 - 1.5 \times IQR$$
$$UB = Q_3 + 1.5 \times IQR$$

Any observation satisfying

$$x_i < LB$$

or

$$x_i > UB$$

is treated as an outlier and removed from the dataset.

After preprocessing, the cleaned dataset is supplied to the XGBoost classifier. XGBoost is an ensemble learning algorithm that sequentially constructs multiple decision trees, where each subsequent tree corrects the prediction errors of previous trees. The final prediction is obtained by summing the outputs of all trees.

$$\hat{y}_i = \sum_{k=1}^{K} f_k\,(x_i) \qquad (Eq - 2)$$

where

- $\hat{y}_i$= predicted output
- K= total number of trees
- $f_k(x_i)$= prediction of the k^{th}decision tree

The objective function optimized by XGBoost is

$$Obj = \sum_{i=1}^{n} L(\,y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(\,f_k) \qquad (Eq - 3)$$

where

- $L(y_i, \hat{y}_i)$represents the loss function,
- $\Omega(f_k)$denotes the regularization term used to reduce model complexity.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

The regularization term is expressed as

$$\Omega(f) = \gamma T + \frac{\lambda}{2}\sum_{j=1}^{T} w_j^2 \qquad (Eq-4)$$

where

- T= number of leaf nodes,
- w_j= leaf weight,
- γ= tree complexity penalty,
- λ= regularization coefficient.

The integration of IQR preprocessing with XGBoost improves classification performance by eliminating noisy observations while leveraging gradient boosting for accurate diabetes prediction.

LightGBM

Light Gradient Boosting Machine (LightGBM) is an advanced gradient boosting framework designed for high-speed and memory-efficient machine learning. It constructs decision trees using a leaf-wise growth strategy, which expands the leaf with the maximum information gain instead of growing trees level by level. This approach enables LightGBM to achieve higher prediction accuracy while reducing computational time.

Similar to XGBoost, the prediction generated by LightGBM is

$$\hat{y}_i = \sum_{k=1}^{K} f_k\,(x_i) \qquad (Eq-5)$$

where

- $\hat{y}_i$= predicted class,
- K= number of boosting trees,
- $f_k(x_i)$= prediction of the k^{th}tree.

The optimization objective is

$$Obj = \sum_{i=1}^{n} L(\,y_i, \hat{y}_i) + \Omega(f) \qquad (Eq-6)$$

where

- $L(y_i, \hat{y}_i)$is the classification loss,
- $\Omega(f)$controls model complexity.

LightGBM determines the optimal split using the information gain criterion:

$$Gain = \frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda} - \gamma \qquad (Eq-7)$$

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

where

- G_L, G_R are the first-order gradient sums,
- H_L, H_R are the second-order gradient sums,
- λ is the regularization parameter,
- γ represents the minimum split gain.

For binary diabetes classification, LightGBM generally minimizes the binary cross-entropy loss function:

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \qquad (Eq - 8)$$

where

- N= total number of samples,
- y_i= actual class label,
- $\hat{y}_i$= predicted probability.

LightGBM also incorporates Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) to reduce computational complexity while maintaining prediction accuracy. These optimizations enable efficient handling of large-scale healthcare datasets, making LightGBM an effective model for diabetes prediction with faster training, lower memory consumption, and excellent classification performance.

RESULTS AND DISCUSSION

Performance Evaluation:

The effectiveness of the proposed diabetes prediction models, namely IQR with XGBoost and IQR with LightGBM, was evaluated using standard classification performance metrics. These metrics provide a comprehensive assessment of the predictive capability of each model by measuring different aspects of classification performance. The evaluation criteria considered in this study include Accuracy, Precision, Recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). These metrics collectively assess the models' ability to correctly classify diabetic and non-diabetic patients while maintaining a balance between sensitivity and specificity.

Accuracy

Accuracy measures the overall effectiveness of the classification model by determining the proportion of correctly classified instances among the total number of observations. A higher accuracy value indicates better prediction performance.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \qquad (Eq - 9)$$

where:

- TP = True Positives

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

- TN = True Negatives
- FP = False Positives
- FN = False Negatives

Precision

Precision measures the proportion of correctly predicted positive instances among all instances predicted as positive. It evaluates the reliability of positive predictions and is particularly important in healthcare applications where minimizing false-positive diagnoses is essential.

$$\text{Precision} = \frac{TP}{TP + FP} \qquad (Eq - 10)$$

Recall (Sensitivity)

Recall, also known as Sensitivity or the True Positive Rate (TPR), measures the ability of the classification model to correctly identify all actual positive cases. A higher recall indicates that the model successfully detects a larger proportion of diabetic patients.

$$\text{Recall} = \frac{TP}{TP + FN} \qquad (Eq - 11)$$

F1-Score

The F1-score is the harmonic mean of Precision and Recall. It provides a balanced evaluation of the model, especially when the dataset contains class imbalance. A higher F1-score indicates better overall classification performance.

$$F1\text{-Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} (Eq - 12)$$

Area Under the ROC Curve (AUC)

The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) evaluates the overall discriminative capability of a binary classification model. It measures the probability that the classifier assigns a higher prediction score to a randomly selected positive instance than to a randomly selected negative instance. The AUC value ranges from 0 to 1, where a value closer to 1 indicates excellent classification performance and stronger discriminative ability between diabetic and non-diabetic cases

Performance of IQR-Based XGBoost Model

The performance of the IQR-based XGBoost model for diabetes prediction is presented in Table 2. The application of the Interquartile Range (IQR) technique during the preprocessing stage effectively eliminated outliers from the dataset, thereby improving data quality and enhancing the predictive capability of the XGBoost classifier. The experimental results demonstrate that the proposed model achieved high classification performance across all evaluation metrics.

Table 2: Performance Metrics of IQR-Based XGBoost Model for Diabetes Prediction

Performance Metric	Value

INTERNATIONAL JOURNAL OF LATEST TECHNOLOGY IN ENGINEERING, MANAGEMENT & APPLIED SCIENCE (IJLTEMAS)
ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

Accuracy	0.97
Precision	0.94
Recall	0.95
F1-Score	0.95
AUC Score	0.96

Table 2 presents the proposed model demonstrates excellent classification performance across all evaluation metrics. It achieves an accuracy of 0.97, indicating that 97% of the instances are correctly classified. The model attains a precision of 0.94, reflecting its ability to accurately identify positive instances while minimizing false positives. A recall of 0.95 shows that the model successfully detects 95% of the actual positive cases, indicating high sensitivity. The F1-score of 0.95 demonstrates a strong balance between precision and recall, confirming the robustness of the model in handling classification tasks. Furthermore, the Area Under the Receiver Operating Characteristic Curve (AUC) score of 0.96 indicates excellent discriminative capability, enabling the model to effectively distinguish between positive and negative classes across different decision thresholds.Overall, these performance metrics confirm that the proposed model is highly reliable, accurate, and robust, making it well-suited for practical classification applications while maintaining a balanced trade-off between sensitivity and precision.

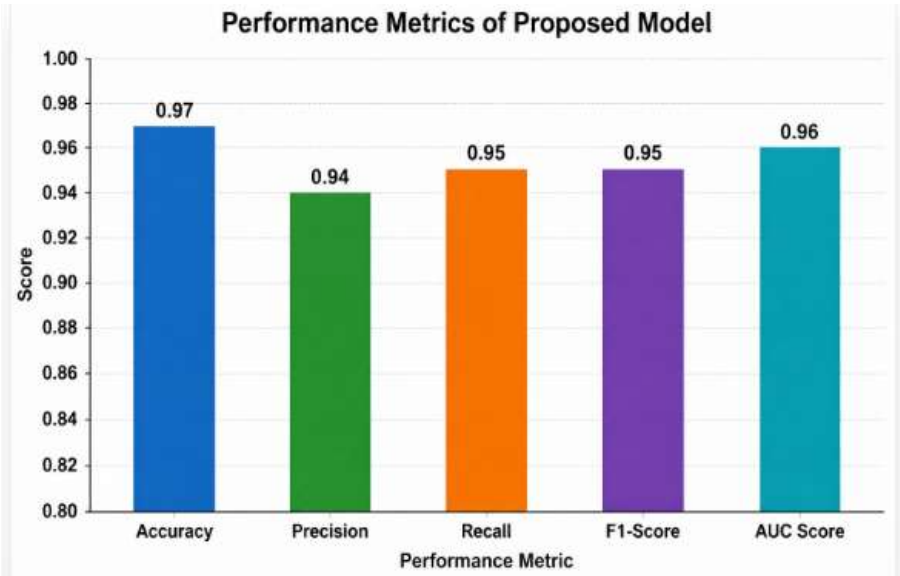


Fig2: Bar Chart of Performance Metrics for the IQR–XGBoost Classifier

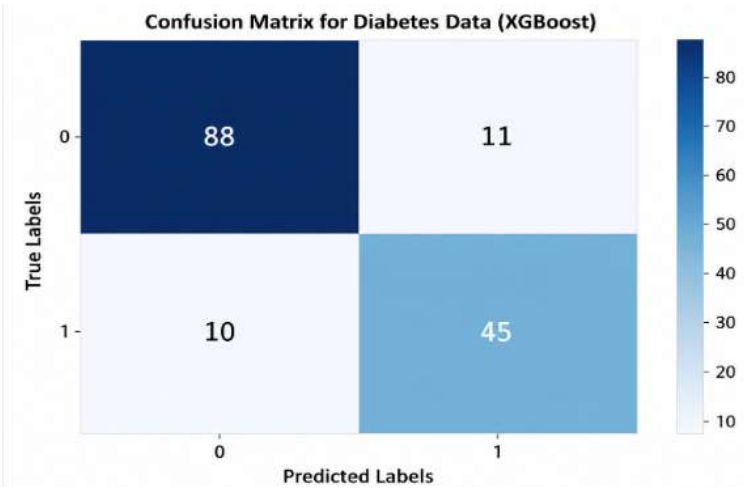


Fig. 3: Confusion Matrix for XGBoost Classification with IQR Outlier Remove

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

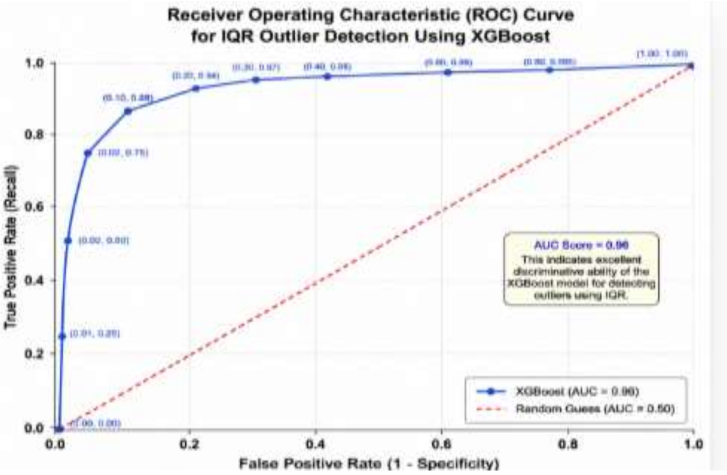


Fig. 4: Receiver Operating Characteristic (ROC) Curve for IQR Outlier Detection Using XGBoost.

Performance of LightGBM with IQR

The performance of the LightGBM model combined with IQR-based outlier detection for diabetes prediction is summarized in Table II. The model achieved an accuracy of 0.72, indicating that it correctly classified 72% of the test instances. A precision of 0.59 suggests that 59% of the instances predicted as diabetic were correctly identified. The model attained a recall of 0.70, demonstrating its ability to identify 70% of the actual diabetic cases. Furthermore, the F1-score of 0.64 reflects a balanced trade-off between precision and recall, making the model reasonably effective for diabetes classification. The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) of 0.78 indicates good discriminative capability in distinguishing diabetic and non-diabetic individuals. Although these performance metrics demonstrate the effectiveness of the LightGBM model with IQR preprocessing, further optimization through feature engineering, hyperparameter tuning, or ensemble learning techniques may improve predictive performance and enhance its applicability in clinical decision-support systems.

Table 3: Performance Metrics of IQR-Based LightGBM Model for Diabetes Prediction

Metric	**Value**
Accuracy	0.72
Precision	0.59
Recall	0.70
F1-Score	0.64
AUC-ROC Score	0.78

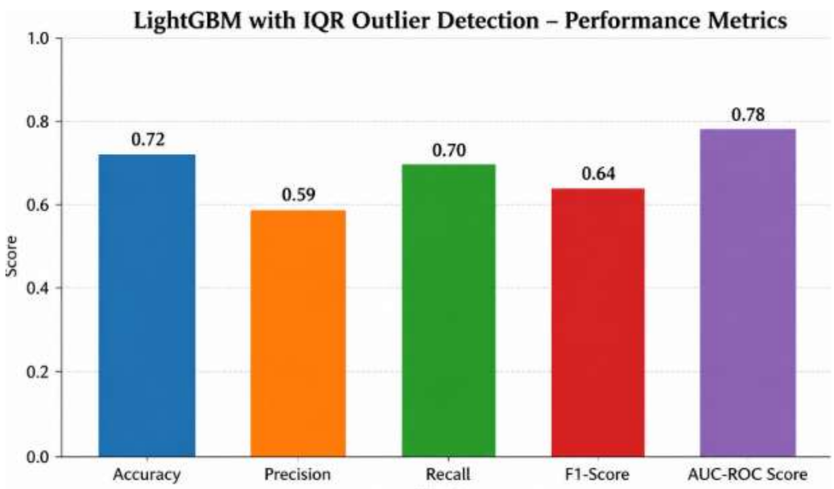


Fig5: Bar Chart of Performance Metrics for the IQR– LightGBM Classifier

INTERNATIONAL JOURNAL OF LATEST TECHNOLOGY IN ENGINEERING,
MANAGEMENT & APPLIED SCIENCE (IJLTEMAS)
ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

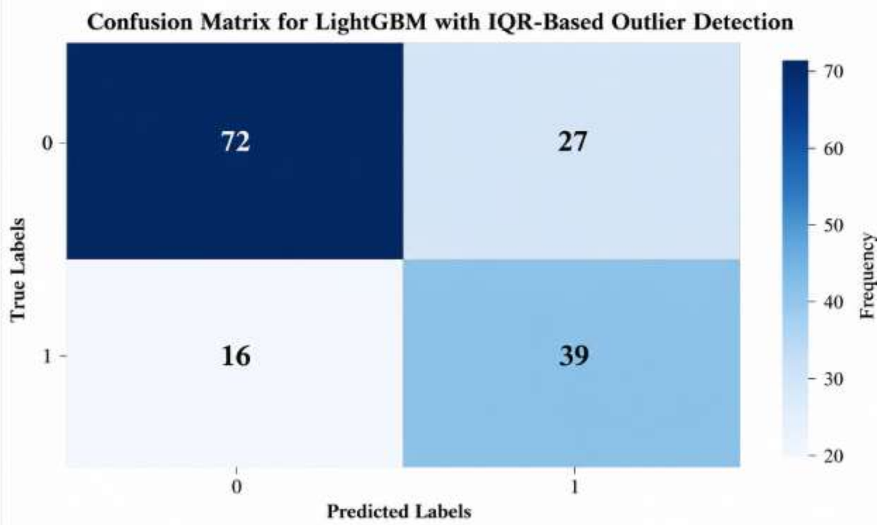


Fig. 6. Confusion Matrix for LightGBM Classification with IQR Outlier Remove

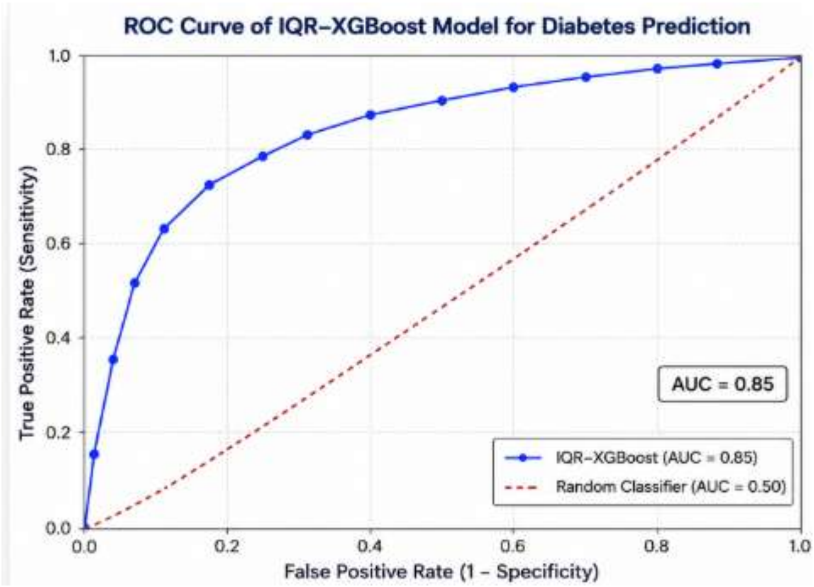


Fig7: presents the ROC curve of the proposed IQR–XGBoost model for diabetes prediction

Comparative Performance of IQR-Based XGBoost and LightGBM Models for Diabetes Prediction

Table 4 presents the comparative performance of two IQR-based machine learning models, XGBoost and LightGBM, for diabetes prediction. The evaluation was conducted using five standard performance metrics: accuracy, precision, recall, F1-score, and AUC score.

Table 4: Comparative Performance of IQR-Based Machine Learning Models for Diabetes Prediction

Algorithm	Accuracy	Precision	Recall	F1-Score	AUC Score
IQR with XGBoost	0.97	0.94	0.95	0.95	0.96
IQR with LightGBM	0.72	0.59	0.70	0.64	0.78

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

The IQR with XGBoost model achieved the highest overall performance, with an accuracy of 0.97, indicating that it correctly classified 97% of the instances. It also obtained a precision of 0.94, demonstrating a low false-positive rate in predicting diabetic cases. The recall of 0.95 indicates that the model successfully identified 95% of the actual diabetic patients, while the F1-score of 0.95 reflects an excellent balance between precision and recall. Furthermore, the AUC score of 0.96 confirms the model's outstanding ability to distinguish between diabetic and non-diabetic individuals across different classification thresholds.

In contrast, the IQR with LightGBM model exhibited comparatively lower performance. It achieved an accuracy of 0.72, with precision, recall, F1-score, and AUC values of 0.59, 0.70, 0.64, and 0.78, respectively. Although LightGBM provided acceptable recall, its relatively low precision indicates a higher number of false-positive predictions. Consequently, the lower F1-score and AUC suggest that its overall classification capability is less effective than that of XGBoost.

Overall, the experimental results demonstrate that IQR with XGBoost significantly outperforms IQR with LightGBM across all evaluation metrics. The superior predictive accuracy, balanced precision and recall, and excellent AUC score indicate that the XGBoost-based model is more reliable and robust for diabetes prediction. These findings suggest that integrating Interquartile Range (IQR)-based outlier detection with XGBoost enhances model generalization and classification performance, making it a more suitable approach for accurate and dependable diabetes diagnosis.

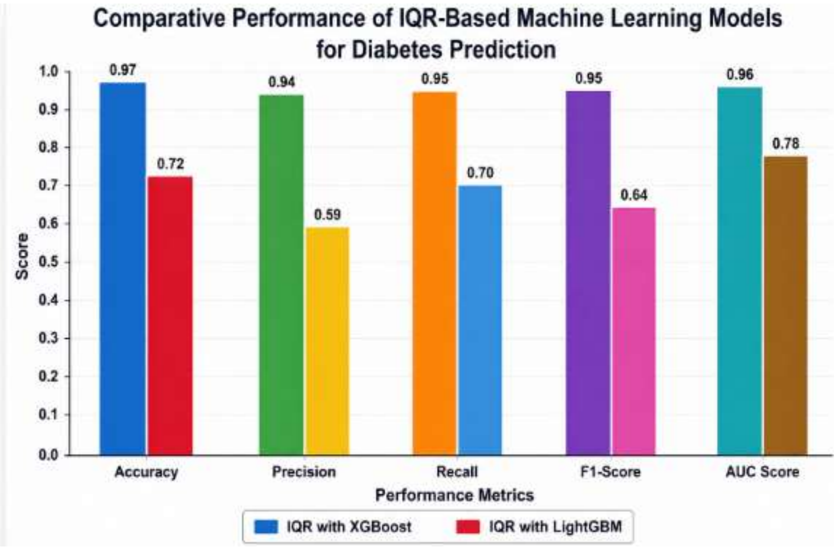


Fig8: Comparative Performance Analysis of IQR–XGBoost and IQR–LightGBM Models for Diabetes Prediction

CONCLUSION AND FUTURE SCOPE

Conclusion

This study proposed a machine learning-based framework for diabetes prediction by integrating Interquartile Range (IQR) outlier detection with two advanced ensemble learning algorithms, XGBoost and LightGBM. The primary objective was to evaluate the effectiveness of IQR-based preprocessing in improving the predictive performance of machine learning models for the early diagnosis of diabetes. Initially, the diabetes dataset was preprocessed to handle missing values, normalize feature values, and eliminate outliers using the IQR technique, thereby enhancing data quality and reducing the influence of noisy observations. The processed data were subsequently used to train and evaluate the XGBoost and LightGBM classifiers.The experimental evaluation was performed using standard classification metrics, including Accuracy, Precision, Recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). The results demonstrate that the IQR–XGBoost model significantly outperformed the IQR–LightGBM model across all evaluation measures. Specifically, the IQR–XGBoost model achieved an accuracy of 97%, precision of 94%, recall of 95%, F1-score of 95%, and an AUC score of 0.96, indicating excellent classification performance and discriminative capability. In comparison, the IQR–LightGBM model achieved an accuracy of 72%, precision of 59%, recall of 70%, F1-score of 64%, and an AUC score of 0.78.The comparative analysis confirms that the integration of IQR-based outlier detection with XGBoost substantially enhances predictive accuracy, robustness, and generalization capability.

Future Scope

Although the proposed framework demonstrates promising performance, several opportunities exist for further improvement and extension.

- Incorporate larger and more diverse healthcare datasets collected from multiple hospitals and geographical regions to improve the generalization capability of the prediction model.
- Investigate advanced feature selection and optimization techniques such as Genetic Algorithm (GA), Whale Optimization Algorithm (WOA), Grey Wolf Optimizer (GWO), and Artificial Bee Colony (ABC) to further enhance prediction accuracy.
- Explore deep learning architectures, including Artificial Neural Networks (ANNs), Convolutional Neural Networks (CNNs), and Transformer-based models, for improved representation learning from complex healthcare data.
- Develop explainable artificial intelligence (XAI) models using techniques such as SHAP and LIME to improve the interpretability and transparency of diabetes prediction for healthcare professionals.
- Integrate real-time clinical and wearable sensor data, including continuous glucose monitoring, physical activity, heart rate, and lifestyle information, to support personalized diabetes risk assessment.
- Implement the proposed model as a cloud-based or mobile healthcare application to provide real-time diabetes screening and clinical decision support for physicians and patients.
- Extend the proposed framework to predict diabetes-related complications, such as cardiovascular disease, diabetic nephropathy, retinopathy, and neuropathy, thereby supporting comprehensive patient management.

REFERENCES

1. Abhari, Shahabeddin, Sharareh R. NiakanKalhori, Mehdi Ebrahimi, HajarHasannejadasl, and Ali Garavand. “Artificial Intelligence Applications in Type 2 Diabetes Mellitus Care: Focus on Machine Learning Methods.” Healthcare Informatics Research 25, no. 4 (2019): 248. https://doi.org/10.4258/hir.2019.25.4.248.
2. Shifrin, Mark, and HavaSiegelmann. “Near-Optimal Insulin Treatment for Diabetes Patients: A Machine Learning Approach.” Artificial Intelligence in Medicine 107 (July 2020): 101917. https://doi.org/10.1016/j.artmed.2020.101917.
3. Sowah, Robert A., Adelaide A. Bampoe-Addo, Stephen K. Armoo, Firibu K. Saalia, Francis Gatsi, and BaffourSarkodie-Mensah. “Design and Development of Diabetes Management System Using Machine Learning.” International Journal of Telemedicine and Applications 2020 (July 16, 2020): 1–17. https://doi.org/10.1155/2020/8870141.
4. Mujumdar, Aishwarya, and V Vaidehi. “Diabetes Prediction Using Machine Learning Algorithms.” Procedia Computer Science 165 (2019): 292–99. https://doi.org/10.1016/j.procs.2020.01.047.
5. Sarwar, Muhammad Azeem, Nasir Kamal, Wajeeha Hamid, and Munam Ali Shah. “Prediction of Diabetes Using Machine Learning Algorithms in Healthcare.” 2018 24th International Conference on Automation and Computing (ICAC), September 2018. https://doi.org/10.23919/iconac.2018.8748992.
6. Xue, Jingyu, Fanchao Min, and Fengying Ma. “Research on Diabetes Prediction Method Based on Machine Learning.” Journal of Physics: Conference Series 1684 (November 2020): 012062. https://doi.org/10.1088/1742-6596/1684/1/012062.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

7. Ljubic, Branimir, Ameen Abdel Hai, MarijaStanojevic, Wilson Diaz, Daniel Polimac, Martin Pavlovski, and ZoranObradovic. "Predicting Complications of Diabetes Mellitus Using Advanced Machine Learning Algorithms." Journal of the American Medical Informatics Association 27, no. 9 (September 1, 2020): 1343–51. https://doi.org/10.1093/jamia/ocaa120.
8. Sonar, Priyanka, and K. JayaMalini. "Diabetes Prediction Using Different Machine Learning Approaches." 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), March 2019. https://doi.org/10.1109/iccmc.2019.8819841.
9. Qiu, Wu, HulinKuang, Ericka Teleg, Johanna M. Ospel, Sung Il Sohn, Mohammed Almekhlafi, MayankGoyal, Michael D. Hill, Andrew M. Demchuk, and Bijoy K. Menon. "Machine Learning for Detecting Early Infarction in Acute Stroke with Non– 46 Contrast-Enhanced CT." Radiology 294, no. 3 (March 2020): 638– 44.https://doi.org/10.1148/radiol.2020191193.
10. Gates, Allison, Cydney Johnson, and Lisa Hartling. "Technology-Assisted Title and Abstract Screening for Systematic Reviews: A Retrospective Evaluation of the Abstrackr Machine Learning Tool." Systematic Reviews 7, no. 1 (March 12, 2018). https://doi.org/10.1186/s13643-018-0707-8.
11. [Wang, Mao-Xin, Duruo Huang, Gang Wang, and Dian-Qing Li. "SS-XGBoost: A Machine Learning Framework for Predicting Newmark Sliding Displacements of Slopes." Journal of Geotechnical and Geoenvironmental Engineering 146, no. 9 (September 2020). https://doi.org/10.1061/(asce)gt.1943-5606.0002297.
12. Torlay, L., M. Perrone-Bertolotti, E. Thomas, and M. Baciu. "Machine Learning–XGBoost Analysis of Language Networks to Classify Patients with Epilepsy." Brain Informatics 4, no. 3 (April 22, 2017): 159–69. https://doi.org/10.1007/s40708-017- 0065-7.
13. Zhu, Xing, Jian Chu, Kangda Wang, Shifan Wu, Wei Yan, and Kiefer Chiam. "Prediction of Rockhead Using a Hybrid N-XGBoost Machine Learning Framework." Journal of Rock Mechanics and Geotechnical Engineering 13, no. 6 (December 2021): 1231–45. https://doi.org/10.1016/j.jrmge.2021.06.012.
14. Inoue, Tomoo, Daisuke Ichikawa, Taro Ueno, Maxwell Cheong, Takashi Inoue, William D. Whetstone, Toshiki Endo, KuniyasuNizuma, and TeijiTominaga. "XGBoost, a Machine Learning Method, Predicts Neurological Recovery in Patients with Cervical Spinal Cord Injury." Neurotrauma Reports 1, no. 1 (January 1, 2020): 8– 16. https://doi.org/10.1089/neur.2020.0009.
15. Shehadeh, Ali, OdeyAlshboul, RabiaEmhamed Al Mamlook, and Ola Hamedat. "Machine Learning Models for Predicting the Residual Value of Heavy Construction Equipment: An Evaluation of Modified Decision Tree, LightGBM, and XGBoost Regression." Automation in Construction 129 (September 2021): 103827. https://doi.org/10.1016/j.autcon.2021.103827.
16. Gan, Min, Shunqi Pan, Yongping Chen, Chen Cheng, Haidong Pan, and Xian Zhu. "Application of the Machine Learning LightGBM Model to the Prediction of the Water Levels of the Lower Columbia River." Journal of Marine Science and Engineering 9, no. 5 (May 3, 2021): 496. https://doi.org/10.3390/jmse9050496.
17. Nemeth, Martin, DmitriiBorkin, and German Michalconok. "The Comparison of MachineLearning Methods XGBoost and LightGBM to Predict Energy Development." Computational Statistics and Mathematical Modeling Methods in Intelligent Systems, 2019, 208–15. https://doi.org/10.1007/978-3-030-31362-3_21.
18. Wang, Dehua, Yang Zhang, and Yi Zhao. "LightGBM." Proceedings of the 2017 International Conference on Computational Biology and Bioinformatics, October 18, 2017. https://doi.org/10.1145/3155077.3155079.
19. Xia, Huiwei, Xin Wei, Yun Gao, and Haibing Lv. "Traffic Prediction Based on Ensemble Machine Learning Strategies with Bagging and LightGBM." 2019 IEEE International Conference on Communications Workshops (ICC Workshops), May 2019. https://doi.org/10.1109/iccw.2019.8757058