International Journal of Computing and Artificial Intelligence

**E-ISSN:** 2707-658X
**P-ISSN:** 2707-6571
Impact Factor (RJIF): 5.57
www.computersciencejournals.com/ijcai
IJCAI 2025; 6(2): 229-237
Received: 17-09-2025
Accepted: 27-10-2025

**Vishal Verma**
Department of Information Technology, Babasaheb Bhimrao Ambedkar University, Lucknow, Uttar Pradesh, India

**Satish Kumar**
Department of Computer Application, Integral University, Lucknow, Uttar Pradesh, India

**Vandna Rani Verma**
CSE Department, Galgotias College of Engineering and Technology, Greater Noida, Uttar Pradesh, India

**Alka Agrawal**
Department of Information Technology, Babasaheb Bhimrao Ambedkar University, Lucknow, Uttar Pradesh, India

**Corresponding Author:**
**Vishal Verma**
Department of Information Technology, Babasaheb Bhimrao Ambedkar University, Lucknow, Uttar Pradesh, India

# An Intelligent Machine Learning Pipeline for Early Diabetes Prediction: CatBoost Ensemble with SMOTTEEN and Optuna Tuning

**Vishal Verma, Satish Kumar, Vandna Rani Verma and Alka Agrawal**

**DOI:** https://doi.org/10.33545/27076571.2025.v6.i2c.202

**Abstract**
Diabetes has been recognized as one of the most widespread diseases worldwide. It arises when the level of glucose in the bloodstream exceeds normal levels. Prediction of diabetes with high accuracy is crucial in the medical industry. Machine learning (ML) techniques play an essential role in building predictive models for healthcare analysis. This study proposes an ensemble approach based on the CatBoost algorithm for early diabetes prediction. To address class imbalance in the dataset, researchers employ the SMOTEENN hybrid sampling technique, and Optuna is utilized for automated hyperparameter tuning. The Diabetes Prediction Dataset (DPD) was preprocessed using data cleaning, IQR-based outlier removal, and label encoding before training. CatBoost was evaluated against several other ML algorithms, including KNN, RF, DT, XGBoost, ETC, LightGBM, and AdaBoost, showing better performance. The proposed hypertuned CatBoost model has shown 98.93% accuracy with better precision, recall, f1-score, and AUC-ROC. In the future, researchers will extend the model to other datasets for generalization and develop predictive models that enable the early detection and forecasting of diabetes progression at the individual patient level by uncovering patterns in clinically captured data.

**Keywords:** Machine learning, Diabetes prediction, CatBoost, SMOTEENN, Optuna

## 1. Introduction
Diabetes is a chronic disease recognized as one of the most widespread diseases worldwide. Glucose obtained from the food consumed serves as the principal energy source for the body. The pancreas, situated adjacent to the stomach, secretes insulin, a hormone that makes it possible for glucose to enter the body's cells. In individuals with this health issue, inadequate insulin production or utilization disrupts this process, resulting in elevated blood glucose levels [1]. Uncontrolled diabetes can result in major health complications, including kidney failure, injury to the nerves, blurry vision, heart attack, and stroke. Effective management through lifestyle changes, medication, and monitoring is important to prevent complications and maintain health [2].

Mainly, diabetes has three types such as Type 1 diabetes (T1D), Type 2 diabetes (T2D), & Gestational diabetes (GD). T1D occurs when the body is unable to generate sufficient insulin, an essential hormone for regulating blood sugar levels. It often begins in adulthood or puberty, requiring lifelong insulin. In T2D, the body is unable to use insulin properly and does not make enough insulin. It is common in adults or young people, but it can occur at any time. It can be controlled through proper exercise, medication, and lifestyle. GD is the most common and often occurs in pregnant women. It usually resolves after the baby is born. This pattern of behaviour raises the likelihood of both the mother and her child developing T2D in the future [3]. Early diagnosis and proper management are important for every type of diabetes. People can prevent themselves from diabetes by following a balanced diet and exercise [4].

Diabetes has a geographically and economically differential burden of morbidity in the world, with the greatest impact in developing countries. According to the International Diabetes Federation (IDF) Diabetes Atlas, 11th Edition (2025), more than 589 million people aged 20-79 years worldwide are living with diabetes. Their number is expected to grow to 643 million in 2030 and 853 million in 2050. In 2021, approximately 6.7 million deaths were linked to diabetes worldwide [5]. With the rising incidence of diabetes

worldwide and the resulting medical costs, early and precise diagnosis has become crucial. Traditional diagnosis techniques often fail to detect diabetes at an early stage, although ML has emerged as a powerful tool to assist in the early detection of diabetes. It can analyze large data as well as find hidden patterns and correlations that may be challenging for physicians to notice [6].

ML yields excellent results in diabetes classification and early detection. Today, a vast amount of data is available about diabetes, including its causes, symptoms, and health effects. Analyzing this data is essential, but not easy, especially when the goal is early diagnosis and treatment. Several ML-based algorithms, including KNN, Random Forest (RF), Decision Tree (DT), ExtraTrees (ETC), XGBoost (XGB), LightGBM (LGBM), and AdaBoost, have been widely used to predict diabetes with reliable accuracy. These models are trained on existing data to identify patterns that can predict the disease early, but each has its own limitations. Despite their advantages, ML methods are often difficult to interpret, limiting their use in clinical settings where reliability and transparency are crucial. Ensemble learning solves this problem by aggregating multiple models to enhance prediction performance. Among ensemble methods, CatBoost, a gradient boosting algorithm, stands out for its performance and high accuracy in handling categorical data, making it suitable for diabetes prediction and similar tasks. [7].

In this study, the researchers propose an ML-based model to accurately identify and predict diabetes. CatBoost is taken as an ensemble algorithm, and the DPD dataset [8] is employed for the experiment. Several preprocessing techniques and the SMOTEENN data sampling technique are used to minimize overfitting of the model. As well as the Optuna technique is used to automate hyperparameter tuning. Proposed models have been evaluated on several metrics, including accuracy, precision, recall, f1-score.

The forthcoming sections are systematically structured as follows: Section 2 examines related work focusing on the application of different ML-based approaches and class balancing techniques. Section 3 is about the methods and materials. Section 4 details and assesses the findings from the experimental study. Finally, Section 5 summarises the studies and suggests directions for future research.

### 1.1 Related Works

Multiple studies have investigated the application of machine learning and deep learning techniques for the accurate identification and prediction of different forms of diabetes. Various data sampling techniques were explored to address class imbalance in the dataset. These approaches leverage advanced computational models to analyse complex relationships among risk parameters. Existing research has employed various ML techniques, which demonstrate their potential to enhance the diagnosis and treatment of diabetes. The primary contributions in this area include the development of a predictive model for diabetes prediction. Also focused on an appropriate feature selection technique for identifying key risk factors. Furthermore, incorporating pre-processing techniques to address issues such as class imbalance and data noise.

P. Suresh Kumar *et al*. [9] developed a CatBoost-based model to predict diabetes at an early stage. They used the Early Stage Diabetes Risk Prediction dataset from the UCI repository that contains 520 instances. They evaluated CatBoost's performance against ML models such as KNN, Multi-Layer Perceptron (MLP), Logistic Regression (LR), Gaussian Naive Bayes (GNB), and Stochastic Gradient Descent (SGD). And CatBoost achieved better results compared to the other models. Saxena *et al*. [10] introduced an optimal feature selection technique to enhance prediction accuracy for diabetes. Correlation Attribute Selection, Information Gain, and PCA feature selection methods were employed on the PIDD to get the best method for feature selection. They tested several classifiers, including MLP, DT, KNN, and RF, and found that RF performed the best.

Ahmed *et al*. [11] developed a framework for diabetes detection, incorporating decision-level fusion. Their hybrid approach combined Support Vector Machine (SVM) and Artificial Neural Network (ANN) models with fuzzy logic, achieving an accuracy of 94.87%. Whig *et al*. [12] introduced an innovative approach for diabetes classification and prediction, leveraging PyCaret, an automated machine learning framework designed to streamline model selection, hyperparameter tuning, and performance evaluation. Dip Das *et al*. [13] presented a paper to investigate the use of ensemble learning methodology towards the prediction of diabetes using the BRFSS Health Indicator dataset. The dataset has 253,680 survey responses for 21 features (history, behaviors, and demographics). They used SMOTEENN resampling technique to deal with the problem of class imbalance. XGBoost is utilized to identify the most relevant features for diabetes prediction. A combination of XGBoost, RF, CatBoost, LightGBM, and KNN classifiers is employed using the Voting Classifier method (Soft Voting). The proposed model achieved 96.40% accuracy, demonstrating strong predictive performance.

Singh Modak *et al*. [14] introduced a diabetic predictive model using LR, SVM, NB, and RF, alongside ensemble methods including XGBoost, LGBM, CatBoost, Adaboost, and Bagging to improve accuracy and robustness by combining predictions from multiple models. Samet Aymaz [15] investigated the optimal data balancing ratios for balancing classes in healthcare data sets for better classification performance. They used data balancing techniques, namely SMOTE, ADASYN, and Borderline-SMOTE. The balancing ratio is further optimized by Particle Swarm Optimization (PSO), Whale Optimization Algorithm (WOA), and Optuna for enhancing the accuracy and efficiency. Performance measures of accuracy, precision, recall, F-score, and resource consumption demonstrate the balance of optimization to improve diagnostic success applicable in healthcare applications.

Julius Olasunmibo Ogunniyi [16] developed a multilingual prediction system of type 2 diabetes that improves both the prediction accuracy and inclusiveness using the CATBoost machine learning algorithm. It combines English and Yoruba phrases and bridges communication on the shortcomings of local health diagnosis. The study is based on a review of 1,197 records from multiple hospitals and a community and evaluates 13 risk factors. Four machine learning algorithms, namely DT, LR, Naive Bayes (NB), and CATBoost, have been tested. Among them, CATBoost performed best with 90.60% and 97.57% accuracy for non-invasive and invasive means. Findings emphasize that the system has the potential for early detection of diabetes, in addition to the need for accessibility and health outcomes, especially in Yoruba-speaking regions.

Hussain and Naaz [17] presented an experimental analysis

evaluating multiple ML and DL models, namely NB, LR, DT, RF, SVM, KNN, and Deep Neural Network (DNN) for diabetes prediction. RF and NB achieved higher accuracy. H. F. Ahmad *et al*. [18] investigated the impact of HbA1c and FPG as input features to predict diabetes. They utilized feature elimination techniques such as feature permutation and hierarchical clustering to enhance the performance of the model. N. Ahmed *et al*. [19] developed a smart web-based application to improve diagnostic accuracy and accessibility. They used a large dataset of Bangladeshi patients for their experiment. SVM was identified as the most accurate approach for diabetes prediction in their study. Khanam and Foo [20] presented a comparative study of ML algorithms for diabetes prediction. Feature reduction methods and WEKA tools were utilized by them. They utilized the PIDD dataset and seven ML algorithms, including NB, DT, KNN, RF, AdaBoost, LR, and SVM.

Mittal Kumar *et al*. [21] proposed an ensemble learning approach for diabetes mellitus classification and prediction with the aid of a soft voting classifier. The better performance of ensemble methods in combining diverse model predictions was demonstrated by their study. Gollapalli *et al*. [22] introduced an innovative stacking ensemble-based learning model for diabetes detection, distinguishing between pre-diabetes, T1DM, and T2DM using a Saudi Arabian dataset. Several classifiers were integrated to enhance prediction accuracy, with the SMOTE technique employed to address class imbalance. And the proposed stacking model achieved 0.9448 of accuracy, 0.9448 of recall, and 0.9470 of precision. Dutta *et al*. [23] conducted an ensemble-based machine learning study for early diabetes prediction, introducing a newly labeled dataset from Bangladesh. Their approach incorporated feature selection, missing data imputation, K-fold cross-validation, and grid search for hyper-tuning to enhance model performance. An accuracy of 73.5% was achieved by the proposed weighted ensemble of DT, RF, XGB, and LGBM. Kaur *et al*. [24] proposed an MCDM-based framework to recommend the most suitable ML technique for diabetes prediction, using methods like WSM, TOPSIS, and VIKOR to rank 10 ML techniques based on eight performance measures. A fusion approach determined final rankings, validated on the PIDD, showing no single ML technique excelled across all measures.

Nipa *et al*. [25] investigated early diabetes prediction using the Sylhet Diabetes Hospital dataset and additional patient records from Bangladesh. Various classifiers, including LR, KNN, SVM, RF, GB, and NN, were applied. Ensemble methods like bagging, boosting, and stacking outperformed individual models. LGBM, stacking, hist gradient boosting, and RF demonstrated stable performance, with SHAP analysis identifying key predictive factors such as age, polyuria, and delayed healing. Chowdhury *et al*. [26] investigated several ML-based models and a data augmentation approach for diabetes detection using the BRFSS dataset. Their study addressed class imbalance challenges, improving the reliability and accuracy of predictive models for medical.

Most researchers have observed several potential challenges, limitations, gaps, and issues in diabetes prediction using machine learning. A primary challenge is class imbalance, where the unequal distribution of data across classes affects model performance [27]. Additionally, a critical problem is remained by feature selection as relevant and high-impact features are significantly influenced by the selection process, affecting prediction accuracy [28]. The limitations observed in existing studies are inclusive of issues like the generalizability of models to diverse patient populations and the prevalence of overfitting.

## 2. Methods and Materials

ML plays a crucial role in diagnosing a disease. In addition, it is helpful to identify the patterns and trends of the disease. In this direction, researchers have used a large diabetes dataset containing 100000 instances to build a robust predictive model. The proposed methodology contains several pre-processing approaches, including SMOTEENN, IQR, and Label Encoding techniques. For the experiment purpose, various ML-based models, including KNN, RF, DT, ETC, XGB, LGBM, AdaBoost, and CatBoost, have been employed. Subsequently, evaluated the efficacy of these algorithms across various evaluation metrics such as accuracy, precision, recall, and f1-score.

### 2.1 Diabetes Dataset

The Diabetes Prediction Dataset (DPD) dataset is used in this study was obtained from the Kaggle open repository [8]. It consists of 100000 instances, including eight input variables and one output variable. The input features are gender, age, hypertension (HT), heart disease (HD), smoking history (SH), body mass index (BMI), HbA1c level (Hb), and blood glucose level (BGL). The target variable is a binary class label indicating whether the individual is diabetic (1) or non-diabetic (0). Among these instances, 8500 belong to the True class, indicating individuals diagnosed with diabetes, while 91500 belong to the False class, representing individuals without diabetes. This class imbalance underscores the need for advanced data balancing techniques such as SMOTEENN to improve model training and performance.

### 2.2 Data Preprocessing

Data preprocessing is crucial as it highly affects the performance and the generalization capabilities of ML models [29]. The key aim of data preprocessing is to eliminate irrelevant data that carries little valuable information and reduce the data modeling biases [30]. In the preprocessing phase, researchers conducted a details analysis to ensure the dataset's quality and suitability for model training. One significant step was deleting 3,854 duplicate rows, which could negatively affect the model's performance. Duplicates might lead to overfitting, so eliminating them was crucial for ensuring the integrity of the training process.

To address the issues with categorical features, we employed Label Encoding. There were two columns, Gender and Smoking History that held categorical values such as 'Male' /'Female' and 'Smoker' /'Non-Smoker'. As machine learning models need numerical data to learn from, we transformed these string values into numeric values by employing Label Encoding. Through this process, a unique number is given to every category so that the model can easily work with the features. Label Encoding is straightforward but useful for binary categories such as these, where the numerical values still capture the inherent meaning of the categories. And the Interquartile Range (IQR) technique was employed to identify and eliminate

outliers. Outliers are values that fall outside the range of normality and may bias the model's performance.
Furthermore, researchers have used SMOTEEN hybrid resampling techniques to address the class imbalance issue.

### 2.3 SMOTE-ENN Technique for Addressing Imbalanced Data

SMOTE-ENN is a combination technique to over-sample the minority class. It's an approach to class imbalance in machine learning, namely when we have a classification task. It formulates two important techniques in its base, which are SMOTE (Synthetic Minority Oversampling Technique) and ENN (Edited Nearest Neighbors). Oversamples the minority class simultaneously and cleans the dataset of ambiguous or noisy instances. The SMOTE creates synthetic instances of the minority class by interpolating between existing minority class samples and their nearest neighbors. This helps to mitigate underrepresentation without merely duplicating samples. On the other hand, the ENN component is a data cleaning technique that removes samples (usually from the majority class) whose class label differs from most of their nearest neighbors, thereby eliminating borderline or mislabeled instances [31]. By applying ENN after SMOTE, the technique refines the data space to enhance the decision boundaries of classifiers.
SMOTE-ENN has been shown to outperform in several domains, such as cancer detection, intrusion detection, and fraud prediction, due to its ability to balance the dataset while removing noisy or overlapping samples. For example, in a comparative study of resampling methods for real-time regression tasks, SMOTE-ENN yielded more robust predictions by reducing variance and improving model generalization across unbalanced datasets. Despite its advantages, SMOTE-ENN may require tuning to avoid excessive removal of valid samples by ENN, especially in high-dimensional or sparse datasets. But when well-calibrated, it provides a powerful method for improving classifier performance on imbalanced datasets [32].

### 2.4 Predictive Models Utilizing Machine Learning for Diabetes Diagnosis

Over the past decade, machine learning has emerged as a pivotal tool in disease detection and prediction, revolutionizing healthcare through data-driven insights and advanced analytical techniques. It is crucial for the identification and prediction of diabetes, as it identifies concealed patterns and trends. Many ML-based models have been proposed to identify a diabetic at an earlier stage, and different results have been reported. In this study, researchers have utilized the most accurate models, namely KNN, RF, DT, ETC, XGB, LGBM, AdaBoost, and CatBoost. A concise explanation of these algorithms is provided in this section of the study.

### 2.5 KNN

The K-Nearest Neighbors (KNN) is simple, non-parametric, lazy supervised machine learning algorithm commonly used for classification and regression problems. It does this by querying the 'K' most similar matches (close neighbors) to a provided query point, employing distance functions such as Euclidean distance. The class (or value) of the query point is the majority class of (average value of) its neighbors. KNN is intuitive and has no training process, but performance may be affected with big data since the prediction is distance based. The choice of 'K' value and distance metric are crucial [33].

### 2.6 Random Forest

Random Forest (RF) is a powerful ensemble learning approach designed for classification tasks. It combines predictions from multiple decision trees to produce a unified outcome. This model utilizes decision trees to sample data rows and columns, optimizing the number of base learners to minimize variance and improve accuracy. RF is renowned for its effectiveness in bagging techniques, playing a pivotal role in reducing prediction errors and enhancing overall model performance within the field [10].

### 2.7 Decision Tree

Decision Tree (DT) is a widely used algorithm for binary classification. It is structured as a hierarchical tree. DT organizes features as branches to predict target values found in the tree leaves. Classification trees are used when the response takes on a finite number of values, showing adaptive partitions that provide class labels, where the leaves consist of these labels. In contrast, regression trees are suited for targets with continuous values, such as real numbers. He Decision Tree can be used with two primary root selection criteria: IG (Information Gain) and Gini. IG is important in the selection of the root node, which is significant in defining the structure and predictive performance of the Decision Tree [34].

### 2.8 Extra Trees Classifier

Extra Trees Classifier (ETC) is an ensemble learning model that creates a number of decision trees and does not prune, using a top-down approach. In contrast to Random Forest, ETC builds the whole training dataset instead of bootstrap replicates. It injects heavy randomization when forming nodes by randomly selecting the attributes and the cut points. This can produce completely random trees, regardless of the training data outputs. The feature and value for the node split of ETC is also randomized, which is one difference from RF. These properties make ETC more robust to overfitting, and it outperforms the state-of-the-art methods on several datasets [35].

### 2.9 XGBoost

XGBoost is one of the powerful ensemble learning methods in machine learning. It is commonly used in supervised regression models due to its ability to improve the accuracy of base learners and optimize objective functions. It employs ensemble learning by integrating predictions of several models when training, which gives the final prediction a unified and more accurate [13].

### 2.10 Light Gradient Boosting Machine

Light Gradient Boosting Machine (LGBM) is a newly introduced powerful ensemble algorithm commonly used for classification and ranking tasks. This employs a histogram-based learning methodology by partitioning the data both vertically and horizontally. This algorithm is valued for its speedy performance, efficiency, and ease of handling massive data. It adopts a tree-based learning algorithm for processing and is ideally suited for high-performance computing and parallel processing [36].

### AdaBoost

AdaBoost stands for Adaptive Boosting, which is an ensemble learning technique that combines multiple weak classifiers to create a strong classifier. In each iteration, AdaBoost focuses more on the sample that are misclassified in previous round, adjusting the weights of training samples accordingly. This adaptive nature helps improve accuracy over time. It can be used in a wide variety of classification and regression tasks.

### CatBoost

CatBoost is a machine learning algorithm developed capable of handling categorical features and utilizes a gradient boosting technique. It relies on ordered boosting and complex regularisation strategies to enhance model accuracy by decreasing overfitting. And like other models, CatBoost takes care of raw categorical variables without the data processing like one-hot encoding. Superior performance and high computational efficiency are shown on data sets with complicated categorical structures, which ensures its stability for classification and regression tasks [13].

### Proposed Model

The proposed model aims to improve the accuracy for early prediction of diabetes through an ensemble model. The Diabetes Prediction Dataset (DPD) from the Kaggle database repository is preprocessed in order to ensure data quality. This includes the dropping of 3,854 duplicate rows. 10,019 anomalies are detected by the IQR method. The Label Encoder technique is used to transform categorical data into numeric. After that, we used the SMOTEENN sampling method to balance the classes. Lastly, the preprocessed dataset was divided in an 80:20 ratio for training and testing the model. The method is developed based on the CatBoost model and fine-tuned to improve its performance. Experiments show that the proposed system effectively enhances the accuracy of disease prediction and represents an effective approach to diabetes prediction. A flowchart of the proposed model is illustrated in Fig. 1. Tables 1 summarize the nature of the dataset.

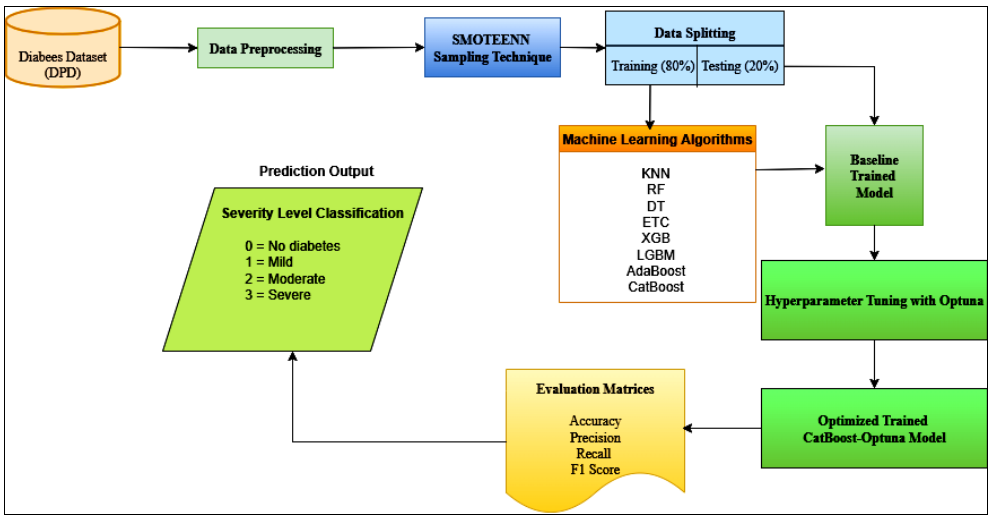


**Fig 1:** Proposed workflow for diabetes prediction using CatBoost-Optuna

**Table 1:** Information about the Dataset Type

| Features | Description | Non-Null Count | Dtype |
|---|---|---|---|
| Gender | Gender (Male/Female) | 100000 non-null | Object |
| Age | Age in years | 100000 non-null | Float64 |
| Hypertension (HT) | An individual has high blood pressure. (1 = Yes, 0 = No) | 100000 non-null | Int64 |
| heart_disease (HD) | An Individual has any heart-related condition. (1 = Yes, 0 = No) | 100000 non-null | Int64 |
| smoking_history (SH) | Describes past or current smoking habits. | 100000 non-null | Object |
| Bmi | BMI is a measure of body fat based on height and weight. | 100000 non-null | Float64 |
| Hba1c_Level (Hb) | Average blood glucose levels over the past 2 or 3 months. | 100000 non-null | Float64 |
| blood_glucose_level (BGL) | Shows the current level of glucose in the blood. | 100000 non-null | Int64 |
| Diabetes | Class variable (0: No diabetes, 1: Diabetes) | 100000 non-null | Int64 |

### Evaluation Matrics

The model has been assessed on the testing dataset using various metrics such as accuracy, precision, recall, and F1 score. The confusion matrix is a fundamental evaluation criterion for the performance of a classifier. There are four possible results: True Positive (TP), the number of positive cases correctly predicted, and True Negative) TN as the number of negative cases accurately described. False Positive (FP), that is, the number of negative samples incorrectly retained within positive ones. Similarly, FalseNegative (FN) is the count of false predictions as negative. And the authors use different measures to assess how well their method performs. The model's classification performance could not be evaluated using these key measures. Limiting criteria are characterized as:

**Accuracy (Acc):** Acc measures how the model's predictions were correct based on the following formula:

$$Acc = \frac{[\text{TruePositive } (TP) + \text{TrueNegative } (TN)]}{[\text{TruePositive } (TP) + \text{TrueNegative } (TN) + \text{FalsePositive } (FP) + \text{FalseNegative } (FN)]} \quad (2)$$

**Recall (Rec):** Rec represents the ratio of actual positive instances correctly classified as positive by a model, calculated as follows:

$$Rec = \frac{TruePositive\ (TP)}{\text{TruePositive } (TP) + \text{FalseNegative } (FN)} \quad (3)$$

**Precision (Prec):** Prec is defined as the ratio of true positive instances to the sum of true positive and false positive instances. It is calculated as follows:

$$Prec = \frac{TP}{\text{TruePositive } (TP) + \text{FalsePositive } (FP)} \quad (4)$$

**F1 Score (F1):** F1 Score is an important metric for unbalanced classes. Accuracy can be misleading. It is, in fact, a harmonic mean of Precision and Recall, which ensures a balanced assessment on the model's performance, calculated as follows:

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

## 3. Results Analysis

This research deals with early-stage diabetes risk prediction, executed in Python 3.8 in Google Colab, and several requisite libraries, such as Pandas, Scikit-learn, and Matplotlib libraries are used. The performance of the model is evaluated based on important metrics such as precision, recall, accuracy, along with f1-score, and AUC-ROC. We have adopted several pre-processing techniques, such as the SMOTEENN hybrid sampling technique, to balance the original dataset and remove outliers through IQR. The proposed CatBoost-Optuna model achieved an AUC-ROC score of 1.0 depicted in Fig. 3, indicating that it discriminates incredibly well between diabetes-positive and diabetes-negative cases. This optimal separation attests to the strength of the model in segregating individuals at risk from non-risk subjects. Furthermore, the model's excellent precision of 99.06% and recall of 98.91% confirm its accuracy of 98.93%, indicating that it is very successful in lowering false positives while simultaneously identifying the majority of actual cases. The model's ideal balance between precision and recall is guaranteed by the F1-score of 98.99%, which is particularly advantageous when working with unbalanced medical datasets.

While other models such as KNN, RF, DT, ETC, XGB, LGBM, and AdaBoost performed competitively, CatBoost outperformed them in terms of precision-recall balance. This shows that CatBoost's improved performance can be ascribed to both its gradient boosting mechanism and the incorporation of Optuna-based hyperparameter optimization, which improved generalization and decreased overfitting. These findings clearly demonstrate the proposed pipeline's potential for real-world use in early diabetes prediction. The detailed experimental results are presented in Table 2, and Fig. 2 shows the graphical representation.

**Table 2:** Classification Report of Machine Learning Models Utilizing SMOTEENN

| Model | Acc | Prec | Rec | F1 | AUC-ROC |
|---|---|---|---|---|---|
| KNN | 0.9742 | 0.9568 | 0.9960 | 0.9760 | 0.9951 |
| RF | 0.9839 | 0.9787 | 0.9910 | 0.9848 | 0.9990 |
| DT | 0.9801 | 0.9785 | 0.9838 | 0.9812 | 0.9798 |
| ETC | 0.9842 | 0.9770 | 0.9935 | 0.9852 | 0.9991 |
| XGB | 0.9802 | 0.9807 | 0.9818 | 0.9813 | 0.9986 |
| LGBM | 0.9779 | 0.9798 | 0.9783 | 0.9791 | 0.9985 |
| AdaBoost | 0.9294 | 0.9194 | 0.9494 | 0.9342 | 0.9828 |
| CatBoost | 0.9856 | 0.9874 | 0.9853 | 0.9864 | 0.9991 |
| Optimized Trained CatBoost-Optuna Model | 98.93 | 99.06 | 98.91 | 98.99 | 1.00 |

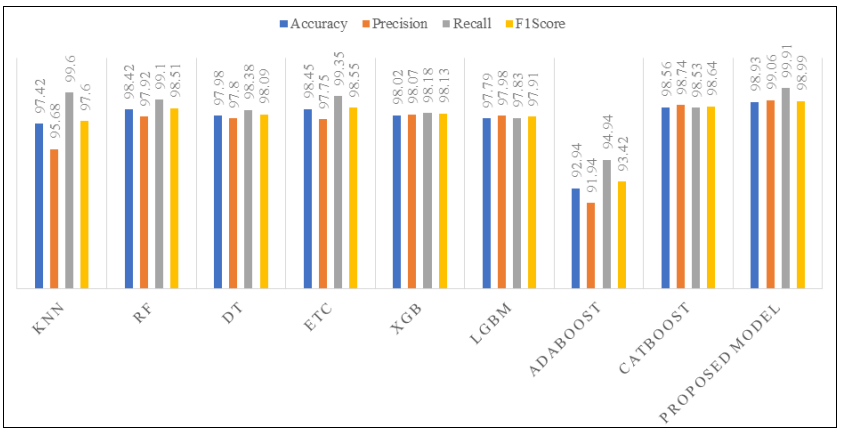


**Fig 2:** Learning Models Results Obtained by Using SMOTEENN

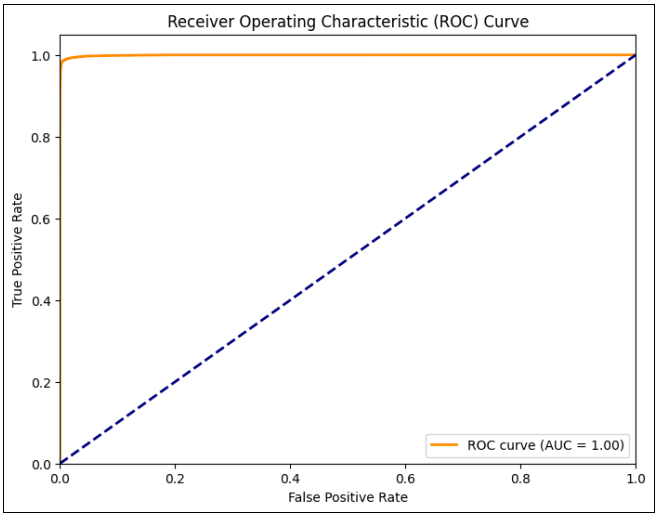


**Fig 3:** AUC-ROC of proposed method

## 4. Discussion and Comparison

A comparative study on ML methods demonstrates the superiority of the proposed CatBoost-Optuna model for early prediction of diabetes as shown in Table 2. Although traditional ensemble methods, RF and the ETC showed good accuracies of 98.39% and 98.42%, respectively, their precision-recall trade-off was slightly weaker compared to CatBoost. This indicates that even if these models succeed in predicting many of the true events, they could be less reliable in minimizing both false positives and negatives.

On the other hand, the proposed CatBoost Optuna model reached an accuracy of 98.93%, and precision, recall, and F1-scores greater than 98.9%. The high precision of the model shows that it is effective in preventing false alarms, while the same high recall indicates its strength in detecting true diabetes cases, as detailed in Fig. 4. The balanced F1-score exhibits CatBoost's performance in both sensitivity and specificity, which is a significant characteristic, especially concerning imbalanced medical data.

The higher accuracy of CatBoost can be explained by two main factors. Its gradient boosting framework naturally addresses overfitting and deals well with categorical features, which traditional models do not. Second, Optuna-based hyperparameter tuning integration enabled the fine-tuning of learning parameters to generalize across datasets [37]. Combined with the SMOTEENN balancing technique, these improvements made it possible for the model to perform best in comparison with other methods, making CatBoost a robust. With its high performance, the model has tremendous potential for clinical use as a decision-support system. With its ability to differentiate diabetic from non-diabetic patients accurately, it can help healthcare providers in the early detection and risk stratification, and timely management, hence lessening the risk of complications and enhancing patient outcomes.

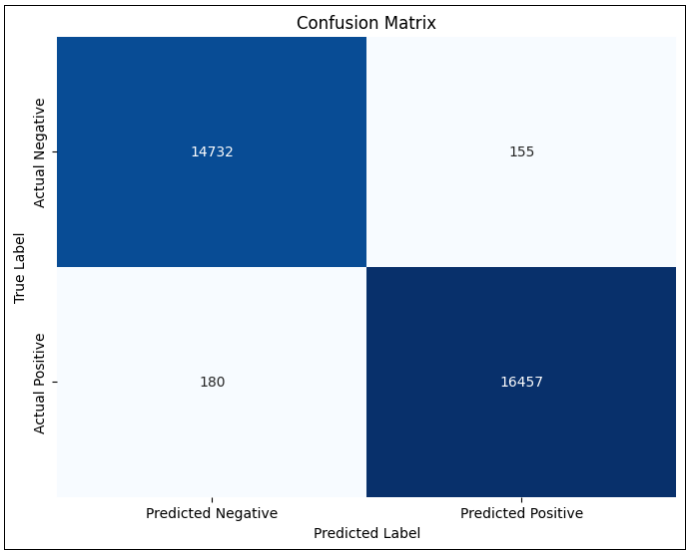


**Fig 4:** The Proposed Model's Confusion Matrix.

### 4.1 Confusion Matrix

The confusion matrix provides a clear visualization of the model's classification performance by showing how well it distinguishes between diabetes-positive and diabetes-negative cases. The performance of the model in terms of AP, AN, PP, and PN is illustrated by the confusion matrix in Fig. 4. In this study, the proposed model correctly identified 14,732 instances as AP and 16,457 instances as AN. However, it misclassified 180 actual positives as negatives (AN) and 130 actual negatives as positives (AP). This indicates that the model not only captures most of the actual diabetic instances but also minimizes false errors, which is

crucial for reliable deployment in healthcare applications.

## 5. Conclusion

Diabetes is now an epidemic worldwide. It's getting worse due to a bad lifestyle, less physical activity, and the environment. The high death rate also shows how important it is to have early diagnosis. This research has resulted in a robust and effective predictive model that integrates hypertuned CatBoost algorithms, along with pre-processing methods, including Label Encoding, SMOTEENN, and IQR, to enhance the accuracy of our predictions. Models are experimented on the DPD dataset (100,000 instances) and measured in terms of ML metrics such as Accuracy, Precision, Recall, F1-score, and AUC-ROC. The CatBoost-Optuna model developed classified more effectively at a rate of 98.93% and an AUC-ROC curve of 1.0, addressing various challenges such as Overfitting and generalizing the model. This advanced model could be a dependable prediction tool to assist clinicians in decision-making.

In the future, researchers will extend the model to other datasets for generalization and develop predictive models that enable the early detection and forecasting of diabetes progression at the individual patient level by uncovering patterns in clinically captured data. This will lead to improved diabetes prevention intervention and management coordination around the world.

## Declarations

**Conflict of Interest:** The corresponding author declares that none of the authors have any conflicts of interest.

## References


1. Westman EC. Type 2 Diabetes Mellitus: A Pathophysiologic Perspective. Front Nutr. 2021;8:1-15. https://doi.org/10.3389/fnut.2021.707371
2. Verma V, Kumar Verma S, Kumar S, Agrawal A, Ahmad Khan R. Diabetes Classification and Prediction Through Integrated SVM-GA. Recent Advances in Computational Intelligence and Cyber Security. 2024;96-105. https://doi.org/10.1201/9781003518587-8
3. Maniruzzaman M, Rahman MJ, Ahammed B, Abedin MM. Classification and prediction of diabetes disease using machine learning paradigm. Health Inf Sci Syst. 2020;8:1-14. https://doi.org/10.1007/S13755-019-0095-Z/TABLES/13
4. Verma V, Kumar S, Agrawal A. Enhancing Early Diabetes Prediction Using Extra Tree Classifier: A Machine Learning Approach. 2025;148-157.
5. International Diabetes Federation (IDF). IDF Diabetes Atlas 2025 | Global Diabetes Data & Insights. 2025. https://diabetesatlas.org/resources/idf-diabetes-atlas-2025/
6. Li Y, Li D, Lin J, Zhou L, Yang W, Yin X, *et al*. Proteomic signatures of type 2 diabetes predict the incidence of coronary heart disease. Cardiovasc Diabetol. 2025;24:1-13. https://doi.org/10.1186/S12933-025-02670-3/FIGURES/4
7. Islam MM, Rifat HR, Shahid MS Bin, Akhter A, Uddin MA, Uddin KMM. Explainable Machine Learning for Efficient Diabetes Prediction Using Hyperparameter Tuning, SHAP Analysis, Partial Dependency, and LIME. Eng Rep. 2025;7:e13080. https://doi.org/10.1002/ENG2.13080
8. Kaggle. Diabetes prediction dataset. Kaggle.com; 2025. http://www.kaggle.com/
9. Kumar PS, Anisha Kumari K, Mohapatra S, Naik B, Nayak J, Mishra M. CatBoost ensemble approach for diabetes risk prediction at early stages. Proc 1st Odisha Int Conf Electr Power Eng Commun Comput Technol (ODICON 2021). 2021;1-8. https://doi.org/10.1109/ODICON50556.2021.9428943
10. Saxena R, Sharma SK, Gupta M, Sampada GC. A Novel Approach for Feature Selection and Classification of Diabetes Mellitus: Machine Learning Methods. Comput Intell Neurosci. 2022;2022:3820360. https://doi.org/10.1155/2022/3820360
11. Ahmed U, Issa GF, Khan MA, Aftab S, Farhan Khan M, Said RAT, *et al*. Prediction of Diabetes Empowered With Fused Machine Learning. IEEE Access. 2022;10:1-12. https://doi.org/10.1109/ACCESS.2022.3142097
12. Whig P, Gupta K, Jiwani N, Hruthika J, Kouser S, Alam N. A novel method for diabetes classification and prediction with Pycaret. Microsyst Technol. 2023;29:1-10. https://doi.org/10.1007/s00542-023-05473-2
13. Das D, Aayushman, Kumar S, Hussain MA, Reddy BR. Diabetes Prediction using Ensemble Learning Techniques. Procedia Comput Sci. 2025;258:3155-3164. https://doi.org/10.1016/J.PROCS.2025.04.573
14. Kumar Singh Modak S, Kumar Jha V, Kumar Jha V. Diabetes prediction model using machine learning techniques. Multimed Tools Appl. 2024;83:38523-38549. https://doi.org/10.1007/s11042-023-16745-4
15. Aymaz S. Unlocking the power of optimized data balancing ratios: a new frontier in tackling imbalanced datasets. J Supercomput. 2025;81:1-62. https://doi.org/10.1007/S11227-025-06919-2/METRICS
16. Ogunniyi JO, Adetunji OJ, Fasanya OI, Fasanya TO, Emuoyibofarhe JO, Olamoyegun MA. A Multilingual Predictive System for Type 2 Diabetes using the CATBoost Machine Learning Algorithm. 2025;1-12. https://doi.org/10.21203/RS.3.RS-6195549/V1
17. Hussain A, Naaz S. Prediction of Diabetes Mellitus: Comparative Study of Various Machine Learning Models. Adv Intell Syst Comput. 2021;1166:103-115. https://doi.org/10.1007/978-981-15-5148-2_10
18. Ahmad HF, Mukhtar H, Alaqail H, Seliaman M, Alhumam A. Investigating Health-Related Features and Their Impact on the Prediction of Diabetes Using Machine Learning. Appl Sci. 2021;11:1173-1185. https://doi.org/10.3390/APP11031173
19. Ahmed N, Ahammed R, Islam MM, Uddin MA, Akhter A, Talukder MA, *et al*. Machine learning based diabetes prediction and development of smart web application. Int J Cogn Comput Eng. 2021;2:229-241. https://doi.org/10.1016/J.IJCCE.2021.12.001
20. Khanam JJ, Foo SY. A comparison of machine learning algorithms for diabetes prediction. ICT Express. 2021;7:432-439. https://doi.org/10.1016/J.ICTE.2021.02.004
21. Kumari S, Kumar D, Mittal M. An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. Int J Cogn Comput Eng. 2021;2:40-46. https://doi.org/10.1016/J.IJCCE.2021.01.001

22. Gollapalli M, Alansari A, Alkhorasani H, Alsubaii M, Sakloua R, Alzahrani R, *et al*. A novel stacking ensemble for detecting three types of diabetes mellitus using a Saudi Arabian dataset: Pre-diabetes, T1DM, and T2DM. Comput Biol Med. 2022;147:105757. https://doi.org/10.1016/J.COMPBIOMED.2022.105757
23. Dutta A, Hasan MK, Ahmad M, Awal MA, Islam MA, Masud M, *et al*. Early Prediction of Diabetes Using an Ensemble of Machine Learning Models. Int J Environ Res Public Health. 2022;19:12378-12388. https://doi.org/10.3390/ijerph191912378
24. Kumar A, Kaur K. A Novel MCDM-Based Framework to Recommend Machine Learning Techniques for Diabetes Prediction. Int J Eng Technol Innov. 2023;20:1-12. https://doi.org/10.46604/ijeti.2023.11837
25. Nipa N, Riyad MH, Satu S, Walliullah, Howlader KC, Moni MA. Clinically adaptable machine learning model to identify early appreciable features of diabetes. Intell Med. 2024;4:22-32. https://doi.org/10.1016/J.IMED.2023.01.003
26. Chowdhury MM, Ayon RS, Hossain MS. An investigation of machine learning algorithms and data augmentation techniques for diabetes diagnosis using class imbalanced BRFSS dataset. Healthc Anal. 2024;5:100297. https://doi.org/10.1016/J.HEALTH.2023.100297
27. Abousaber I, Abdallah HF, El-Ghaish H. Robust predictive framework for diabetes classification using optimized machine learning on imbalanced datasets. Front Artif Intell. 2024;7:1499530. https://doi.org/10.3389/FRAI.2024.1499530/BIBTEX
28. Natarajan K, Baskaran D, Kamalanathan S. An adaptive ensemble feature selection technique for model-agnostic diabetes prediction. Sci Rep. 2025;15:1-12. https://doi.org/10.1038/S41598-025-91282-8
29. Karuppasamy M, Rani JM, Poorani K. Metaheuristic Feature Selection for Diabetes Prediction with P-G-S Approach. Procedia Comput Sci. 2025;252:165-171. https://doi.org/10.1016/J.PROCS.2024.12.018
30. Nithya B, Asha V, Sreeja SP, Amit Yadav V, Aditya M, Rai A. A Data-Driven Approach for Early Detection and Prevention of Diabetes Mellitus Using Machine Learning. Proc Int Conf Vis Anal Data Vis (ICVADV 2025). 2025;44-49. https://doi.org/10.1109/ICVADV63329.2025.10961906
31. Ashraf I, Ur Rehman A, Adimabua Ojugo A, Diwakar M, Rani R, Jaiswal G, *et al*. Enhancing liver disease diagnosis with hybrid SMOTE-ENN balanced machine learning models—an empirical analysis of Indian patient liver disease datasets. Front Med (Lausanne). 2025;12:1502749. https://doi.org/10.3389/FMED.2025.1502749
32. B G, P J, G K, M IN, Mohanarathinam, Velusamy J. Optimized SMOTE for Imbalanced Data Handling in Machine Learning. Proc 3rd Int Conf Adv Comput Comput Technol (InCACCT 2025). 2025;349-354. https://doi.org/10.1109/INCACCT65424.2025.11011340
33. Sivaraman M, Sumitha J. An Efficiency of KNN and Decision Tree Techniques for Diabetes Prediction. Proc 6th Int Conf Electron Commun Aerosp Technol (ICECA 2022). 2022;1418-1424. https://doi.org/10.1109/ICECA55336.2022.10009233
34. Mujumdar A, Vaidehi V. Diabetes Prediction using Machine Learning Algorithms. Procedia Comput Sci. 2019;165:292-299. https://doi.org/10.1016/J.PROCS.2020.01.047
35. Saeed H, Ahmed M. Diabetes type 2 classification using machine learning algorithms with up-sampling technique. J Electr Syst Inf Technol. 2023;10:1-10. https://doi.org/10.1186/S43067-023-00074-5
36. Rufo DD, Debelee TG, Ibenthal A, Negera WG. Diagnosis of Diabetes Mellitus Using Gradient Boosting Machine (LightGBM). Diagnostics. 2021;11:1714-1725. https://doi.org/10.3390/DIAGNOSTICS11091714
37. Daliya VK, Ramesh TK. A Cloud-Based Optimized Ensemble Model for Risk Prediction of Diabetic Progression—An Azure Machine Learning Perspective. IEEE Access. 2025;13:11560-11575. https://doi.org/10.1109/ACCESS.2025.3528033